Estimating the usefulness of distorted natural images using an image contour degradation measure

David Rouse, Sheila Hemami, Romuald Pépion, Patrick Le Callet

Estimating the usefulness of distorted natural images using an image contour degradation measure

David M. Rouse,[1,*] Sheila S. Hemami,[1] Romuald Pépion,[2] and Patrick Le Callet[2]

[1]*Visual Communications Laboratory, School of Electrical and Computer Engineering, Cornell University, 356 Rhodes Hall, Ithaca, New York 14850, USA*

[2]*IRCCyN, Université de Nantes, Rue Christian Pauc, 44306 Nantes, France*

**Corresponding author: dmr58@cornell.edu*

Received July 20, 2010; revised November 5, 2010; accepted November 8, 2010; posted November 19, 2010 (Doc. ID 131986); published January 24, 2011

Quality estimators aspire to quantify the perceptual resemblance, but not the usefulness, of a distorted image when compared to a reference natural image. However, humans can successfully accomplish tasks (e.g., object identification) using visibly distorted images that are not necessarily of high quality. A suite of novel subjective experiments reveals that quality does not accurately predict utility (i.e., usefulness). Thus, even accurate quality estimators cannot accurately estimate utility. In the absence of utility estimators, leading quality estimators are assessed as both quality and utility estimators and dismantled to understand those image characteristics that distinguish utility from quality. A newly proposed utility estimator demonstrates that a measure of contour degradation is sufficient to accurately estimate utility and is argued to be compatible with shape-based theories of object perception.

OCIS codes: 110.2960, 110.3000, 110.3925.

1. INTRODUCTION

Imaging systems that capture, process, compress, transmit, and/or store natural images [1] supply information to humans to permit or to facilitate the performance of a particular task. For instance, people working in the public safety sector (e.g., law enforcement, fire control, and emergency services) use natural imaging systems in real-time scenarios to make immediate decisions about how best to respond to an incident [2,3]. In another example, investigators not only examine recordings obtained with video surveillance systems, but also introduce such recordings as evidence for criminal investigations [4–6].

Consumer imaging systems (e.g., digital cameras) directly used by human observers to perform a particular task capture a broad class of source content and are vulnerable to a broad class of distortions, including compression and transmission errors. When operating with limited resources (e.g., communication bandwidth or memory storage), such imaging systems can produce visibly distorted natural images. A visibly distorted image could impede a human's ability to perform a task and provoke inappropriate responses, or it could have no impact at all. Understanding the impact of distortions is clearly important to system designers, users, as well as the subjects who may be captured. Poorer task performance implies that the distorted image is less useful to a human observer than its undistorted counterpart: the "perceived utility" decreases. The perceived utility characterizes the usefulness of a distorted image as a surrogate for a reference (i.e., undistorted) natural image. For such systems and the images generated by them, an objective estimator of perceived utility would facilitate current and future system design, optimization, and improvement.

Prior work on the perceived utility of natural images can be traced back to the Boston University Optical Research Laboratory formed in 1946, where the human viewing the images was first studied as a component in a reconnaissance imaging system [7]. Later, Johnson quantified task performance in terms of empirically determined sampling criteria for detection, recognition, and identification of a target object [8,9]. The sampling criteria were specified in terms of the number of resolved cycles along the minimum dimension of the target object and established the level of object discrimination with respect to the distance of the target object. Johnson's criteria provide basic guidelines for the design of imaging sensors and the expected performance for a given task (i.e., target recognition).

Other work has investigated alternatives and refinements to Johnson's criteria [10–12]. For example, recognition of a target has been demonstrated to be equivalent to the detection of an equally sized circular disk, which allows for imaging devices to be characterized in terms of the smallest detectable circular disk [10]. A recent study observed that Johnson's criteria was restricted to the objects used in Johnson's study [11]. In another example, Vollmerhausen *et al.* proposed a targeting task performance (TTP) metric that accounts for variations among imaging sensors and computes the integral of the square root of the product of the target contrast, the sensor frequency response, and the contrast sensitivity function of the human visual system (HVS) [12]. The TTP metric was demonstrated to predict task performance more accurately than Johnson's criteria [12].

The impact of various image compression artifacts on task performance has been investigated. One study investigated the use of uncompressed and compressed synthetic aperture radar imagery captured by an airborne sensor to perform various tasks (e.g., vehicle counting and vehicle classification) and reported the relationship between task performance and the compression ratio [13]. Given the same compression

ratio, Irvine *et al.* observed that wavelet-based compression techniques yield better task performance than standard JPEG compression [13]. Another study conducted a target identification experiment using uncompressed and compressed close-range thermal imagery containing one of a finite number of known targets [14]. O'Shea *et al.* demonstrated that the TTP metric can be used to predict task performance of compressed imagery using the frequency response of a parameterized Gaussian blur as the sensor frequency response in the TTP metric, where the parameters of the Gaussian blur were selected to fit the experimental results [14].

A fundamental limitation of the prior work on image utility is the use of *a priori* knowledge about the target objects imaged. The experiments conducted to measure task performance train observers to identify a specific set of targets that will appear in the test images [12,14] or prompt observers to perform specific tasks that imply information about the potential content of the image (e.g., vehicle counting) [13]. The models developed in the prior work also incorporate *a priori* knowledge about the target object(s) such as the contrast of the target [12,14]. Practical use of such *a priori* knowledge in models requires (1) a mechanism that correctly associates known target information with the image under evaluation, which increases the complexity of the model, and (2) a database of target information, which limits the scope of images to which the model can be reliably applied. In short, the results from prior work are tailored to specific applications and provide little insight into the underlying image characteristics that allow human observers to achieve a desired task performance level for a broad class of images, and the work in this paper seeks to understand and identify those underlying image characteristics.

Over the past three decades, consumer imaging systems have been largely studied in the context of perceived quality to characterize the perceptual resemblance of a distorted image to a reference (either known or implied) [15–24]. Objective estimators of perceived quality have been proposed that are designed according to various principles (e.g., signal fidelity measures or HVS models), and these estimators are then tuned to or trained on image databases containing distorted images with subjective scores. Such image databases contain distortions typically affecting consumer imaging systems; for example, the LIVE and CSIQ image databases [25,26] contain images with distortions due to blur, compression, transmission errors, additive noise, and/or global contrast loss. Thus, such estimators are expected to accommodate a broad class of source content and distortions, and various estimators have achieved very good predictive performance of perceived quality for these databases.

The work presented in this paper is motivated by the prior work in both image quality and utility and expands the previous narrowly studied definitions of utility in a manner that allows both a broader evaluation of utility as well as a characterization of the underlying image characteristics that impact usefulness. Unlike the specific tasks performed with images in prior work, the "task" is instead to report the content of an image as it is gradually improved from an initially extremely distorted and unrecognizable version to a visually lossless [27] version. A novel suite of experiments presented here provides utility scores for distorted images, and quality scores are collected using a standard test methodology. Distortions were strategically selected to disrupt various spatial frequencies in a broader sense than those traditionally studied in perceived quality experiments.

An analysis of the resulting relationship between perceived quality and perceived utility demonstrates that an image's perceived quality does not imply that image's usefulness and vice versa. Therefore, an objective estimator that accurately estimates perceived quality scores cannot accurately estimate perceived utility scores and vice versa. These results motivate a thorough analysis of the images to understand the image characteristics that produce distorted but useful images for human observers. We assess the performance of several objective estimators as both quality and utility estimators. Although most of these objective estimators have been designed to estimate perceived quality, they serve as signal analysis tools not only to develop an understanding of those image characteristics that impact usefulness but also to suggest signal analysis tools for an objective utility estimator.

Two objective estimators are shown to accurately estimate utility. The first is an objective estimator that is customarily used as a quality estimator. A modified version of this estimator, in which the modifications adjust the relative importance of distortions across spatial frequencies to the overall objective estimate, is shown to generate the most accurate estimates of perceived quality among the objective estimators evaluated.

The second objective estimator is the newly proposed natural image contour evaluation (NICE) utility estimator, which was inspired by the importance of contour information to the HVS for object perception [28–30]. NICE is based on the hypothesis that degradations to image contours restrict the content that an image conveys to a human and decrease perceived utility. In particular, NICE estimates utility as a function of both lost and introduced contour information in a distorted image when compared with a reference image.

To the best of our knowledge, no experimental methods exist to measure the perceived utility of distorted natural images when the task is to report the content of an image. This paper reports the first usage of such experimental methods as well as a subsequent analysis. Section 2 presents the proposed experimental methodology used to collect perceived utility scores. Several standard methods are available to collect perceived quality scores for distorted natural images, and Section 3 reviews the experimental methodology we used to collect perceived quality scores. Experimental results illustrating the relationship between the perceived utility and perceived quality scores are presented in Section 4. Section 5 reviews objective estimators that are assessed as both utility and quality estimators of distorted natural images in Section 6. The results from both the subjective experiments and the analysis of objective estimators as utility and quality estimators are discussed in Section 7. General conclusions are provided in Section 8.

2. METHODS: PERCEIVED UTILITY SCORES

A distorted natural image is viewed as a surrogate for an undistorted, reference image. A perceived utility score quantifies the usefulness of that distorted image with respect to the reference image for a task. More useful images provide more information about the image content to an human.

Two meaningful anchors on the perceived utility scale describe the usefulness of an image: the recognition threshold (RT) equivalence class and the reference equivalence class (REC). The RT equivalence class, henceforth denoted the RT, specifies an equivalence class of maximally degraded images from which humans accurately recognize the "basic content" of the reference image. The perceived utility score of the RT can distinguish useful distorted images from useless distorted images formed from a reference image. In particular, an image with a perceived utility score greater than that of its RT is useful, whereas an image with a perceived utility score less than that of the RT is useless. Humans recognize at least the basic content of useful images but recognize nothing in useless images.

The basic content of a reference image is subjective for our task, which is reporting the content of an image, so a specific experiment (see Subsection 2.C.3) was conducted to estimate the RT. In that experiment, observers read descriptions (provided by anonymous observers) of distorted images and judged if the description indicated that the "writer" recognized the basic content of the reference image. This allowed the collective responses from all the observers define the basic content of the reference image.

The REC specifies an equivalence class of images, including the reference image, that yield the same interpretation of the content as the reference image. Images in the REC may contain signal degradations that may or may not be visible to a human observer but still convey the same information as the reference image. A visually lossless image could contain signal distortions, yet remain visually indistinguishable from the reference image, so a visually lossless image belongs to the REC. Any distorted image whose perceived utility score is statistically equivalent to that of a visually lossless image formed from the same reference image belongs to the REC.

Two experiments [31] were conducted to obtain perceived utility scores. The first experiment acquires subjective data that were processed (see Subsection 2.D) to produce relative perceived utility scores for a collection of distorted natural images generated from each reference image. These relative perceived utility scores correspond to a unique range of values that only are meaningful for distorted images formed from a specific reference image. The relative perceived utility scores for the RT and the REC of each reference image are used to linearly map the relative perceived utility scores to a common range of values. On this common range of values, the RT is indicated by a perceived utility score of 0, and the REC is indicated by a perceived utility score of 100. The subjective data obtained in the second experiment is used to estimate the RT of each reference image. The REC did not need to be estimated from experimental data because both the reference image and any visually lossless image belong to the REC. A visually lossless image generated via JPEG-2000 (J2K) compression using the dynamic contrast-based quantization (DCQ) strategy [32] defined the REC of each reference image (see Subsection 2.A.3).

The remainder of this section describes the methods used to collect subjective data and produce perceived utility scores. First, the distortion types used to construct reference/distortion image sequences are described. Then, the methods are reported for the experiments conducted using these sequences to acquire subjective data to (1) produce relative perceived utility scores and (2) estimate the RTs of reference images. Last, the derivation of perceived utility scores from the collected subjective data is explained.

A. Reference/Distortion Image Sequences

Sequences of decreasingly distorted natural images were generated from a reference natural image. Each sequence corresponds to a specific distortion and evolves such that subsequent images in the sequence gradually refine detail or information relative to the previous images. For brevity, such a sequence is henceforth denoted (1) generically as a reference/distortion sequence and (2) more specifically by explicitly indicating either the reference image name, the distortion, or both (e.g., reference/JPEG denotes a sequence of JPEG distorted images corresponding to the same undisclosed reference). The reference/distortion sequences were formed by varying a single parameter that controlled the level of distortion. For a single reference subjected to a single distortion, perceived utility is assumed to exhibit a monotonically, nondecreasing relationship with decreasing distortion level. Thus, as a reference/distortion sequence evolves toward a visually lossless image, the perceived utility does not decrease. The sequences of distorted images that correspond to different distortions served as test stimuli in the experiments. Select images from the airplane/J2K + DCQ sequence are shown in Fig. 3.

Each distortion is spatially correlated with the reference natural image and disrupts different image characteristics. The image characteristics disrupted include the spatial frequency content, contour integrity (i.e., edges), and the level of detail (i.e., textures). Example images with each distortion are shown in Fig. 1, and Table 1 summarizes each distortion. Subsections 2.A.1, 2.A.2, 2.A.3, 2.A.4, and 2.A.5 describe the five distortions evaluated in the experiments.

1. JPEG: Quantized Discrete Cosine Transform (DCT) Coefficients

JPEG achieves lossy compression of natural images by quantizing block-based DCT coefficients [33]. The quantization strategy implemented in the source code library provided by the Independent JPEG Group [34] is used and parameterized by $P_{\text{jpeg}} \in [0, 100]$, which scales the example luminance component quantization table suggested in the JPEG specification [35]. A sequence of images with JPEG compression artifacts evolves by increasing the parameter P_{jpeg}.

2. BLOCK: Extreme Blocking Artifacts

Extremely low-rate JPEG images effectively replace each 8×8 block of pixels with their average value. To simulate this, a reference/BLOCK sequence of images has extreme blocking artifacts and evolves by decreasing the quantization step-size Q_{avg} of the average block pixel value.

3. J2K + DCQ: Quantized Discrete Wavelet Transform Coefficients

The lossy J2K image compression standard represents natural images as a linear combination of wavelet basis functions [36]. Distortions are introduced by quantizing the basis function coefficients found using a discrete wavelet transform to achieve a desired encoding bitrate, R. The DCQ strategy assigns quantization step sizes according to a measure of visual

(a) *airplane* (reference) (b) JPEG compression (c) BLOCK: Extreme blocking artifacts

(d) J2K+DCQ: JPEG-2000 compression using the DCQ strategy (e) TS: Texture smoothing (f) TS+HPF: TS plus high-pass filtering

Fig. 1. Original reference airplane image and distorted images illustrating the five distortions described in Subsection 2.A. The JPEG and BLOCK distortions are introduced by quantizing coefficients of a block-based DCT. J2K + DCQ distortions result from quantizing coefficients of a discrete wavelet transform according to the DCQ strategy [32]. TS distortions are induced via TV regularization to smooth texture regions with limited disruption to edges. A HPF that removes low-frequency signal information from images with TS distortions produces the TS + HPF distortions. Table 1 contains descriptions of each of the distortions.

distortion parameterized by characteristics of the image, the wavelet subband coefficients, and the display. The DCQ strategy's visual distortion measure distinguishes visually lossless images from visibly distorted images, so the DCQ strategy can specify subband quantization step sizes for lossy compression that yield a visually lossless image. A reference/J2K + DCQ sequence of images has distortions due to J2K compression using the DCQ strategy and evolves by increasing the encoding bitrate, R.

4. Texture Smoothing (TS)

Edges distinguish objects and regions (i.e., sky and rooftop) in natural images that convey substantial meaning to human observers, whereas textures generally provide secondary information about these objects or regions. Furthermore, the extrastriate visual cortex exhibits the greatest response to images that retain contour information and lack texture information [30]. The apparent significance of edges to the HVS inspired the evaluation of distortions that deliberately smooth texture regions in images with limited disruption to edges.

Total variation (TV) regularization traditionally has been used to remove noise from images by producing piecewise smooth images that lack textures [37]. TV regularization executed via soft thresholding of undecimated Haar wavelet coefficients in all subbands, except the low-frequency residual subband, smooths texture regions in natural images [37–40]. A five-level undecimated Haar wavelet transform is used. A reference/TS sequence of images has distortions due to TS and evolves by decreasing a smoothing parameter γ that controls the degree of TS induced by soft thresholding.

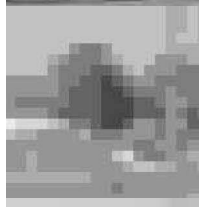
5. TS + HPF: TS plus High-Pass Filtering

Low-frequency content is not critical to preserve the appearance of edges, which commonly coincide with object boundaries in natural images, so images subjected to TS and high-pass filtering were evaluated. When viewing high-pass filtered images, observers necessarily cannot use very low-frequency content by squinting, moving, or otherwise blurring the appearance of the stimulus to interpret the image content. A high-pass filter (HPF) that removes low-frequency content from images with TS distortions produces the TS + HPF distortions.

B. Experiment 1: Subjective Data to Derive Relative Perceived Utility Scores

This experiment collected subjective data that was processed to derive relative perceived utility scores of distorted images formed from the same reference image. Distorted images of the same reference image but subjected to different distortions were compared using a paired comparison test methodology. The images compared were selected from reference/distortion sequences corresponding to the same reference

Table 1. Summary of Image Distortions Studied[a]

Distortion	Description	Parameter Versus Distortion Level	Example	Magnified Example
None	Reference airplane image	N/A		
JPEG	Quantized DCT coefficients according to the lossy JPEG image compression standard. Parameterized by JPEG quality parameter P_{jpeg}.	Increasing P_{jpeg} decreases the level of distortion.		
J2K + DCQ	Quantized discrete wavelet transform coefficients using quantization step-sizes specified by the DCQ strategy for a target encoding bitrate, R.	Increasing R decreases the level of distortion.		
BLOCK	Replace each 8×8 block of pixels by their average and quantize this average pixel value using the quantization parameter Q_{avg}.	Decreasing Q_{avg} decreases the level of distortion.		
TS	TS with limited disruption to image edges. Parameterize by TS parameter γ.	Decreasing γ decreases the level of distortion.		
TS + HPF	TS (i.e., TS distortions) plus high-pass filtering. Parameterize by TS parameter γ.	Decreasing γ decreases the level of distortion.		

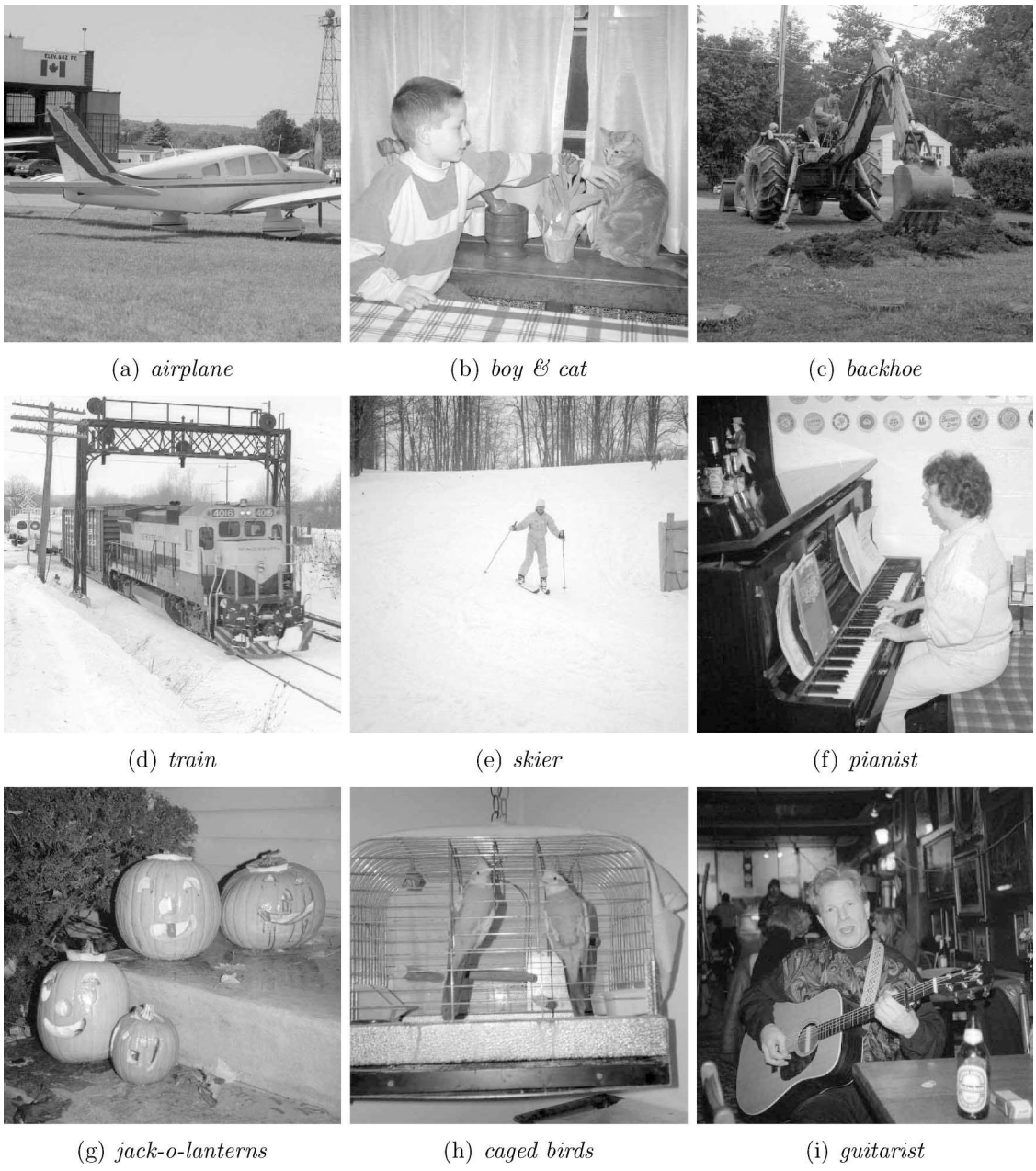
[a]The relationship between the distortion parameter and the level of distortion is described for each distortion. For a reference image subjected to one distortion type, utility and quality are assumed to exhibit a monotonically, nondecreasing relationship with decreasing distortion level.

image but different distortions. The comparisons of images with different distortions were used to align different reference/distortion sequences for the same reference image. For example, these comparisons allow the images from both an airplane/J2K + DCQ sequence and an airplane/TS sequence to be placed in relation to one another in terms of their relative perceived utility. For the same reference image, all reference/distortion sequences corresponding to each distortion were aligned, and these aligned sequences can be merged to form a single sequence of increasingly useful images that contain all distorted images of the same reference image.

1. Stimuli

Nine grayscale natural images of size 512×512 pixels were cropped from original natural images and served as the reference images for these experiments. The content of the natural images consisted of either one or two main objects (e.g., an airplane or a boy and a cat) or a human in action (e.g., skiing or playing guitar). The nine natural images used in the experiments are shown in Figs. 1(a) and 2.

A collection of distorted images was formed by selecting a broad range of distortion levels from each reference/distortion sequence corresponding to each reference image and distortion. Specifically, images with JPEG distortions were formed using JPEG parameter values $P_{\text{jpeg}} = 1$, 2, 5, 10, 20, and 50. Images with BLOCK distortions were formed using quantization step sizes $Q_{\text{avg}} = 400$, 200, and 1. Six images with J2K + DCQ distortions were formed using encoding bitrates logarithmically equally spaced from $R = 0.01$ to R_{VL}, where R_{VL} denotes the bitrate of a visually lossless image formed



(a) *airplane* (b) *boy & cat* (c) *backhoe*

(d) *train* (e) *skier* (f) *pianist*

(g) *jack-o-lanterns* (h) *caged birds* (i) *guitarist*

Fig. 2. Natural images serving as reference images for the experiments.

using the DCQ strategy and J2K compression. Four of the six images from the airplane/J2K + DCQ sequence are shown in Fig. 3. Images with TS and TS + HPF distortions were formed using smoothing parameters $\gamma = 2048$, 446, 97, 21, 5, and 1. The entire collection contained 243 distorted images.

2. Procedure

A paired comparison testing methodology was used to collect subjective responses. Soft copies of the distorted images were presented on a display at a distance of approximately four picture heights. Observers were asked to select an image from a pair of distorted images corresponding to the same reference image in response to the query "Which image tells you more about the content?" Most of the observers were Francophones, and for those observers, the query was presented in French as "Quelle est l'image qui donne le plus d'information sur le contenu de l'image?" The distorted images in each pair correspond to the same reference image but different distortions (e.g., airplane with J2K + DCQ distortions and airplane with TS + HPF distortions). Each observer provided responses for a pair of images once. Certain pair comparisons were determined to be unnecessary based on responses collected in a preliminary experiment (e.g., comparing the most distorted image with J2K + DCQ distortions to the least distorted image with TS distortions), so the number of comparisons for each reference image was reduced. The images in each pair were simultaneously presented side by side on the display, and the placement of the pair of images on the display was randomized. The order that pairs were presented to observers was randomized.

Because of the large number of comparisons, the paired comparison tests were split into four testing sessions.

(a) $R = 0.01$, $U = -57$, $Q = 1.1$ (b) $R = 0.03$, $U = -14$, $Q = 1.2$

(c) $R = 0.08$, $U = 15$, $Q = 1.8$ (d) $R = 0.23$, $U = 46$, $Q = 2.5$

Fig. 3. Four images from the airplane/J2K + DCQ sequence used in Experiment 1 (Subsection 2.B). J2K + DCQ distorted images are parameterized using the encoding bitrate R in bits per pixel (see Table 1). The encoding bitrate of the visually lossless airplane image specified by the DCQ strategy is $R_{VL} = 1.85$ bits/pixel. The perceived utility (U) scores and perceived quality (Q) scores obtained via the subjective experiments are provided for each image.

Observers completed each session in approximately 30 min. Distorted images corresponding to the reference images airplane, boy and cat, caged birds, guitarist, and train were compared in the first two test sessions. J2K + DCQ, TS, and TS + HPF distorted images were included in the first session, and JPEG, BLOCK, TS, and TS + HPF distorted images were included in the second session. Both TS and TS + HPF distorted images appear in both sessions, so that the combined responses from each session also can be used to determine the relationship among J2K + DCQ distorted images and both BLOCK and JPEG distorted images via transitivity.

Distorted images corresponding to the reference images backhoe, jackolanterns, pianist, and skier were compared in the last two test sessions. The last two sessions were designed such that observers compared half of the distorted images in a single test session, and the distorted images in each session spanned the full range of distortion levels tested. All five types of distortions appeared in each of these last two test sessions.

3. Observers

A total of 82 observers with verbally verified normal or corrected-to-normal acuity participated in the experiment over the four test sessions. Forty naive, Francophone observers participated in the first test session. An analysis of the results obtained from the first test session revealed that fewer observers would yield statistically equivalent results, so the remaining test sessions were conducted with fewer observers. In the second test session, ten naive, Francophone observers and ten expert, French- or English-speaking observers participated. Twenty-two naive, Francophone observers participated in the last two sessions with 11 observers per session.

C. Experiment 2: RTs of Natural Images

The experiment to estimate RTs for each of the nine reference images subjected to J2K + DCQ, TS, and TS + HPF distortions consisted of two parts. In the first part, observers called *writers* provided descriptions of the distorted images. In the second part, new observers called *readers* read these descriptions and decided which description indicated that the writer recognized the image content. Since writers typed their descriptions, response time is not a suitable indicator of recognition. The experimental methods used to estimate the RTs of the nine reference images are described.

1. Stimuli

To accurately estimate observer RTs of the reference images, reference/distortion sequences were constructed for each reference image using a dense set of distortion parameters for the J2K + DCQ, TS, or TS + HPF distortions. Reference/J2K + DCQ sequences contained 20 images corresponding to encoding bitrates R that were logarithmically equally spaced from 0.01 to 0.30 bits/pixel. The choice of extremely low bitrates guaranteed that unrecognizable images appear at the beginning of the sequence. Both reference/TS and reference/TS + HPF sequences contained 24 images corresponding to smoothing parameters γ that were logarithmically equally spaced from 2048 to 1. The first image of a reference/TS sequence contained an image with only very low-frequency content, and the first image of a reference/TS + HPF sequence contained an image with a constant valued, gray image. With nine reference images and three distortions, there were a total of 27 reference/distortion sequences.

2. Part 1: Procedure to Collect Descriptions of Distorted Natural Images

In this part of the experiment, which is similar in design to that of Bruner and Potter [41], observers called *writers* viewed a distorted image and typed a brief description of the recognizable image content. The images that a writer viewed and described were ordered such that a writer cycled through each image of one reference/distortion sequence in order of decreasing distortion level. After completely viewing one reference/distortion sequence, the writer cycled through a new reference/distortion sequence corresponding to a different reference image and possibly a different distortion.

A writer necessarily viewed and described the images of at most nine reference/distortion sequences, each sequence corresponding to a different reference image. The order that the reference/distortion sequences were presented to each writer was randomized. Participants completed this task in about 30 min.

3. Part 2: Procedure to Identify RTs from Descriptions Collected in Part 1

In this part of the experiment, observers called *readers* who had not previously viewed the images read the descriptions produced by the writers.

This experiment consisted of consecutive trials. In each trial, a reader read all the descriptions provided by an unidentified writer for the images of a single reference/distortion

sequence. The reference image corresponding to a reference/distortion sequence was simultaneously presented to the reader to compare with the descriptions, but information about the distortion viewed by the writer was hidden from the reader. The list of descriptions typed by a writer were ordered for the reader such that the first description corresponded to the first image of the reference/distortion sequence (i.e., an unrecognizable image), and the last description corresponded to the last image of the sequence. In each trial, the reader was instructed to select the first description that indicated the basic content of the reference natural image had been recognized. Trials were randomized for each reader.

This experiment was split into four sessions to alleviate observer fatigue. No time limit was imposed, and observers completed each session in approximately 30 min.

4. Observers

A total of 49 observers with verbally verified normal or corrected-to-normal acuity participated in the experiments to estimate RTs for the nine reference images. Forty-six English-speaking observers (i.e., writers) participated in the experiment that collected descriptions of images in sequences corresponding to the different distortions. Nine to 13 observers viewed and described the distorted images in the reference/J2K + DCQ sequences for all nine reference images. Not all observers viewed a reference/J2K + DCQ sequence of images corresponding to each of the nine reference images. Twelve observers viewed and described the distorted images in the reference/TS and reference/TS + HPF sequences for all nine reference images. Three English-speaking observers (i.e., readers) participated in the experiment to identify RTs from writers' descriptions.

D. Perceived Utility Scores from Subjective Data

Perceived utility scores were obtained using the subjective data acquired in the two experiments described in Subsections 2.B and 2.C. The process to obtain perceived utility scores is described as three steps.

1. Relative Perceived Utility Scores from Subjective Data

Relative perceived utility scores were derived from the subjective data collected using the paired comparison test method (see Subsection 2.B). In particular, given two differently distorted images formed from the same reference image, the subjective data collected for the pair of images was used to estimate the actual probability that one distorted image is more useful to a human than the other.

Bradley and Terry specified a mathematical model that relates the probability that the response to stimulus X_i is greater than the response to stimulus X_j to a continuum of raw scale values that ranks the collection of stimuli $\{X_i\}_{i=1}^{n}$ according to some measure of merit [42]. This mathematical model was used to derive relative perceived utility scores (i.e., the raw scale values). For a reference image X_{ref}, let X_i denote a distorted image formed from X_{ref}, and let p_{ij} denote the probability that image X_i conveys more information to a human about the content of X_{ref} than image X_j. The Bradley–Terry model was used to map the estimates of p_{ij}, based on the subjective data, to relative perceived utility scores.

Distorted images subjected to the same distortion were not compared in the paired comparison test because perceived utility is assumed to exhibit a monotonically, nondecreasing relationship as the distortion level decreased in the reference/distortion sequences. This assumption was imposed by explicitly defining the estimate of the probability p_{ij} for two types of comparisons. First, for comparisons of an image with itself, the estimate of p_{ii} was set to 0.5, since observers were expected to choose either image with equal probability. Second, for two different distorted images corresponding to the same reference/distortion sequence, the image with less distortion was assumed to have greater perceived utility than the image with more distortion. This second assumption was imposed by setting $p_{ij} = 0.99$ when image X_i and X_j belong to the same reference/distortion sequence (e.g., a JPEG distortion sequence), but the level of distortion for X_i is less than that of X_j. The images used in the paired comparison test were broadly spaced in terms of the distortion level to accommodate this second assumption. For example, suppose X_{R_1} and X_{R_2} are two J2K + DCQ distorted images formed from the reference image using encoding bitrates R_1 and R_2, where $R_1 < R_2$. Because a larger encoding bitrate implies a lower level of distortion for J2K + DCQ distortions, the second assumption was imposed by setting $P(X_{R_2} > X_{R_1}) = 0.99$.

For each reference image, relative perceived utility scores for the corresponding set of distorted images were obtained from the estimates of p_{ij} using a generalized linear model, which Critchlow and Flinger demonstrated is equivalent to the maximum-likelihood method used by Bradley and Terry [43]. The estimates of p_{ij} were either generated from the subjective data or explicitly defined to impose the assumptions regarding the relationship among perceived utility and the distortion parameters for a single distortion. In addition to producing relative perceived utility scores, this data provides a mapping from each distortion parameter to the relative perceived utility scores for each reference image, which was used in the next step.

2. Relative Perceived Utility Scores for the RT and the REC

The RT and the REC of each reference image are used as anchors to map the relative perceived utility scores to the common utility scale (see Subsection 2.D.3). The estimates of the relative perceived utility scores for the RT and REC are described.

The subjective data from the second experiment (see Subsection 2.C) were used to estimate the relative perceived utility score coinciding with the RT of each reference image. The processed subjective data from the first experiment was used to construct mappings from each distortion parameter to the relative perceived utility scores. The RT for each reference/distortion sequence was estimated in terms of the corresponding distortion parameter based on the results from the experiments described in Subsection 2.C (e.g., the RT for a J2K + DCQ sequence was specified in terms of the encoding bitrate R). The relative perceived utility score of the reference/distortion sequence's RT was found by linear interpolation using the mappings from each distortion parameter to the relative perceived utility scores. For a reference image, this yields several estimates of the relative perceived utility score for the RT, one corresponding to each distortion. The relative perceived utility score for the actual RT is estimated as the average of the relative perceived utility scores for the RT

for each distortion because the relative perceived utility scores for the RT for each distortion were found to be statistically equivalent.

Both the reference image and any visually lossless image belong to the REC. Thus, the relative perceived utility score coinciding with the minimum bitrate visually lossless image generated via J2K compression using the DCQ strategy was used to define the relative perceived utility score of the REC (see Subsection 2.A.3). These visually lossless images were included in the paired comparison experiments, so the relative perceived utility scores of the REC of each reference image were directly estimated.

3. Perceived Utility Scores: Relative Perceived Utility Scores Mapped to a Common Utility Scale

Perceived utility scores were obtained by mapping the relative perceived utility scores to a common utility scale, where the RT was mapped to a perceived utility score of 0 and the REC was mapped to a perceived utility score of 100. The relative perceived utility scores for the RT and the REC were used to define a linear mapping from relative perceived utility scores for the distorted images generated from the same reference image to perceived utility scores on the common utility scale.

3. METHODS: PERCEIVED QUALITY SCORES

Human judgments of perceived quality generally indicate the perceptual resemblance of an image to a reference and are quantified by a perceived quality score. The reference is either (1) an explicit, external natural image that is presented to the observer or (2) an internal reference based upon observer expectations that is only accessible to the observer. Despite the vagueness of the term "quality," observers frequently attend to particular distortions (e.g., "blocky," "blurry," "sharp," etc.) to draw conclusions about the perceived quality [44].

Distorted natural images have been studied more often in the context of perceived quality than perceived utility, and several objective estimators have been developed to estimate perceived quality (see Section 5). The relationship between perceived quality and perceived utility is unclear; however, a poor quality image is expected to be less useful than an excellent quality image. If perceived quality accurately estimates perceived utility, then existing objective quality estimators should be suitable as utility estimators. Otherwise, those image characteristics that differentiate judgments of perceived quality from those of perceived utility need to be determined to properly design both quality and utility estimators robust to a variety of distortions.

An experiment was conducted to acquire perceived quality scores for the same images for which perceived utility scores were obtained to understand the relationship between quality and utility. The methods employed to acquire perceived quality scores are reported.

A. Stimuli

The nine reference images and the 243 distorted images formed from these reference images according to the methods described in Subsection 2.B.1 served as test stimuli in this experiment.

B. Procedure

The absolute category rating (ACR) [45] testing methodology [46] was used to collect perceived quality opinions of distorted images from human observers and consists of consecutive trials. In each trial, an observer was presented with a stimulus for 10 s. Then, the display was set to a constant gray background, and the observer was immediately requested to provide a opinion score that indicated his perceived quality of the previously displayed stimulus. The reference images were included in the test stimuli evaluated by the observer, and an observer was unaware if a stimulus was a distorted or reference image. The order of the stimuli presented was random and varied for each observer.

A discrete category rating scale was used that has five categories. Observers provide opinions of quality using the adjectives "bad," "poor," "fair," "good," and "excellent" that define the quality categories. The observers participating in the experiment were Francophones; the rating scale respectively translated to French is "mauvais," "médiocre", "assez bon," "bon," and "excellent."

To alleviate observer fatigue due to prolonged evaluation sessions, the test was split into two sessions, each containing roughly half of the stimuli. Observers completed each session in approximately 30 min and rested for 5 min between the two testing sessions.

C. Observers

Twenty-six naive, Francophone observers with verbally verified normal or corrected-to-normal acuity participated in the experiment, and one observer was rejected as an outlier according to criteria specified in the VQEG multimedia phase I report [47]. The 25 opinion scores from the remaining 25 observers were used to produce perceived quality scores for each stimulus.

D. Perceived Quality Scores from Subjective Data

Observers provided quality judgements that correspond to one of the five category levels (i.e., "bad," "poor," "fair," "good," and "excellent"). These five levels were mapped to the integers on the range 1 to 5 and yield observer opinion scores. The perceived quality score [48] for each test image was computed by averaging the corresponding observer opinion scores.

4. RESULTS: QUALITY IS NOT A PROXY FOR UTILITY

The subjective data collected in Sections 2 and 3 provide perceived utility scores and perceived quality scores for a collection of distorted natural images. An analysis of the resulting relationship between the perceived quality scores and the perceived utility scores is reported and followed by a summary of the image characteristics that appear to influence human judgments of quality and utility, respectively, based on an analysis of the distortions. Example images that illustrate that quality is not a proxy for utility are then presented and discussed.

A. Relationship between Quality and Utility

Perceived quality scores lie on the closed interval $\mathcal{Q} = [1, 5]$, whereas perceived utility scores lie on the set of real numbers $\mathbb{R}$ with 0 denoting the RT and 100 denoting the REC. Images with perceived utility scores less than 0 are unrecognizable

and useless, and images with perceived utility scores greater than 100 are more useful than the reference image.

The relationship between quality and utility was analyzed only for those images whose perceived utility scores lie on the closed interval $\mathcal{U} = [-15, 115]$. No images had perceived utility scores greater than 115, but many images ($n = 80$) had perceived utility scores less than −15. Differences between perceived utility scores for images well below the RT convey less information about utility, since these values result from comparisons of two unrecognizable images. Furthermore, unrecognizable images were rated as having "bad" quality: the perceived quality scores for these images have small standard deviation and both mean and median approximately equal to 1 [49]. Images whose perceived utility scores fall just below the RT were included because Bruner and Potter reported that human observers, especially adults, tend to maintain incorrect hypotheses about the actual content when viewing reference/distortion sequences beginning with a very distorted, unrecognizable images as compared to observers that first view a reference/distorted sequence beginning with a less distorted image [41]. Our experiments to estimate RTs had observers first view very distorted unrecognizable images in the reference/distortion sequences, so including images whose perceived utility scores lie on the interval $[-15, 0]$ accounts for possible overestimates of the RTs due to the phenomenon reported by Bruner and Potter.

To test whether quality is a robust proxy for utility, both correlation and accuracy statistics were used. Specifically, quality is not a robust proxy for utility if (1) perceived quality scores and perceived utility scores are weakly correlated and (2) perceived quality scores inaccurately estimate perceived utility scores. The Pearson linear correlation r, the Spearman rank correlation ρ, and the Kendall rank correlation τ are used to quantify the relationship between perceived quality scores and perceived utility scores [50]. The rank correlation measures, the ρ and τ, quantify the discrepancies between the rank order of the two sets of subjective scores. Neither ρ nor τ are affected by a monotonic, nonlinear mapping.

The root mean squared error (RMSE) and the outlier ratio (OR) were chosen to quantify the accuracy with which perceived quality scores estimate perceived utility scores. The RMSE was computed after fitting the perceived quality scores and the perceived utility scores to a monotonic, nonlinear mapping [see Eq. (1)]. The OR is the proportion of nonlinearly mapped quality scores (i.e., the utility score estimated from quality) that lie outside the 95% confidence interval of the perceived utility score.

Monotonic nonlinear functions were fitted to the subjective scores and used to map perceived quality scores to the utility range, since perceived quality exhibits a nonlinear relationship with perceived utility (see Figure 4). Let $\mathcal{Q} = [1, 5]$ denote the domain of the quality range, and let $\mathcal{U} = [-15, 115]$ denote the domain of the utility range. Let q_i and u_i respectively denote the perceived quality score and perceived utility score of image i. The nonlinear function $f\colon \mathcal{Q} \to \mathcal{U}$ given as

$$f(q) = a\log(q) + b \tag{1}$$

maps perceived quality scores to the utility range, and the parameters $\{a, b\}$ were found by minimizing the sum of the squared error based on the residuals $\{f(q_i) - u_i\}_{i=1}^{n}$, where n is the number of images with both perceived quality and perceived utility scores. The fit was considered sufficient if the residuals exhibit a Gaussian distribution. The Jarque–Bera (JB) normality test determines if a collection values come from an unspecified Gaussian distribution [51], was applied to the set of residuals $\{f(q_i) - u_i\}_{i=1}^{n}$, and concluded that they did come from an unspecified Gaussian distribution at the 95% confidence level.

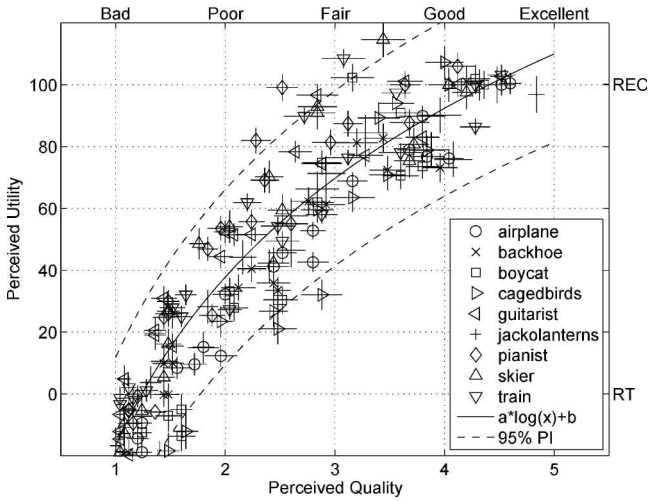


Fig. 4. Quality is not a suitable proxy for utility. The scatterplot shows the relationship between perceived utility scores and the perceived quality scores for nine reference images. The symbols indicate the reference image corresponding to each subjective score. The RT and the REC are denoted on the axis corresponding to perceived utility scores. The quality adjectives are denoted on the axis corresponding to the perceived quality scores. Standard error bars have been included for both subjective scores. In each figure, the fitted nonlinear mapping from the abscissa to the ordinate is denoted by the solid curve, and the 95% PI for the fitted nonlinear mapping is denoted by the dashed curves. See also Fig. 5.

The two scatterplots in Figs. 4 and 5 illustrate the nonlinear relationship between quality and utility for the nine reference images and five distortions with perceived utility indicated on the left ordinate. In each scatter plot, the quality adjectives delineating the quality rating scale have been provided on the top abscissa, and the two anchors, the RT and the

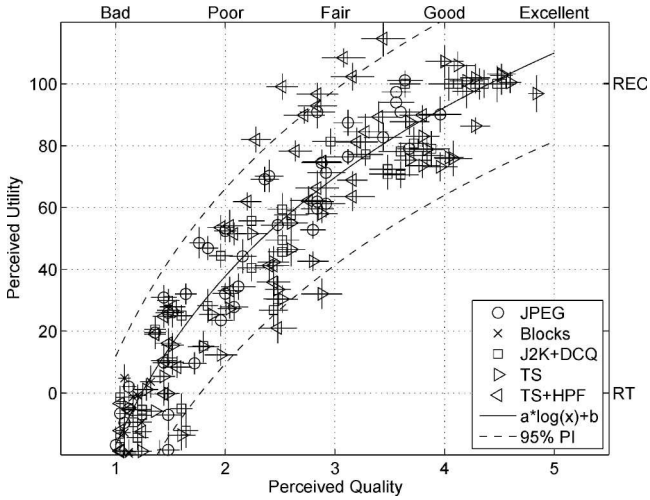


Fig. 5. Perceived utility versus perceived quality where the symbols indicate the distortion (cf. Figure 1) corresponding to each subjective score. See caption of Fig. 4.

REC, associated with perceived utility are indicated on the right ordinate. The symbols in Figs. 4 and 5 distinguish subjective scores according to the reference image and the distortion, respectively. The solid curve in each figure corresponds to the fitted nonlinear mapping from the abscissa to the ordinate [i.e., Eq. (1)], and the dashed curves define the 95% prediction interval (PI) for the fitted nonlinear mapping.

The nonlinear relationship between utility and quality indicates that the quality of a test image generally does not accurately predict its usefulness. The slope of the nonlinear relationship between utility and quality is positive and decreases with increasing quality, which indicates that variations in quality correspond to smaller variations in utility as quality increases. For example, there are test images rated as having perceived quality ranging from "fair" to "excellent" that have high perceived utility.

The relationship between quality and utility was analyzed for the entire collection of distorted images as well as subsets of the collection that were formed by treating (1) quality, (2) distortion type, and (3) reference image (i.e., scene content) as factors. The quality range spans the interval $[1,5]$, and three "levels" of the quality factor were defined for analysis: low quality $[1,2.25)$, medium quality $[2.25,3.75]$, and high quality $(3.75,5]$. Subsets of distorted images spanning these different regions of quality were analyzed because the distorted images used in the experiment span a wide range of distortion levels ranging from unrecognizable to visually lossless. The five distortion types correspond to the "levels" of the distortion type factor: JPEG, BLOCKS, J2K + DCQ, TS, and TS + HPF. Subsets of distorted images corresponding to different distortion types were analyzed because each distortion type disrupts different image characteristics. Subsets of distorted images corresponding to different reference images were analyzed because different image characteristics may affect the relationship between quality and utility for each scene.

Statistical differences in either correlation or accuracy among the different levels of a factor (i.e., quality region or distortion type) preclude a reliable predictive relationship between perceived quality and perceived utility. Statistical differences between two correlation values were determined using a z test after applying the Fisher transformation to the correlation values [52,53]. Statistical differences between accuracy statistics were identified by analyzing the squared errors $\{(f(q_i)-u_i)^2\}_{i=1}^{n}$ using a one-way analysis of variance (ANOVA) to determine if any of the mean squared errors (MSEs) statistically differ for a particular factor [53]. If ANOVA indicated that the accuracy differed according to a particular factor, then Tukey's multiple comparison procedure was used to identify which levels (e.g., high quality or J2K + DCQ) of that factor had statistically different MSEs. The comparison results are reported as p values, where p values greater than 0.05 indicate that at the 95% confidence level the MSEs differ among the two levels of the factor that are compared. The OR is a binomial random variable, and statistical differences between two OR values are determined via a z test at the 95% confidence level using the Gaussian approximation of a binomial random variable [53].

Table 2 summarizes the correlation and accuracy statistics for all images and subsets of distorted images when either the quality region or the distortion is considered as a factor. The monotonic, nonlinear mapping [i.e., Eq. (1)] affects the Pearson linear correlation between the subjective scores. The Pearson linear correlation computed before applying the nonlinearity is denoted r, and it is denoted r_{fit} when computed

Table 2. Results Summarizing the Relationship between Perceived Quality and Perceived Utility[a]

Factor	Image Subset	n	r	ρ	τ	RMSE	r_{fit}	OR
	All	163	0.909	0.919	0.750	14.2	0.925	0.58
Quality region	Low quality	72	**0.819**	**0.791**	**0.606**	12.4	**0.812**	0.58
	Medium quality	63	0.620	0.625	0.458	**17.3**	0.627	0.67
	High quality	28	0.603	0.583	0.402	8.7	0.614	0.32
Distortion	JPEG	39	**0.931**	**0.938**	**0.795**	11.2	**0.939**	0.62
	BLOCKS	6	0.228	0.116	0.138	6.3	0.221	0.00
	J2K + DCQ	42	**0.953**	**0.953**	**0.825**	11.5	**0.955**	0.45
	TS	38	**0.963**	**0.934**	**0.769**	11.0	**0.957**	0.50
	TS + HPF	38	0.884	0.868	**0.690**	**16.5**	0.894	0.71
Reference image	Airplane (set 1)	18	**0.981**	**0.976**	**0.905**	6.0	**0.986**	0.28
	Backhoe (set 1)	16	**0.968**	**0.945**	**0.812**	7.7	**0.972**	**0.31**
	Guitarist (set 1)	21	**0.940**	**0.966**	**0.865**	8.5	**0.977**	**0.43**
	Jackolanterns (set 1)	18	**0.953**	**0.975**	**0.892**	7.4	**0.974**	0.22
	Boy and cat (set 2)	16	**0.936**	0.895	0.740	12.9	**0.949**	**0.56**
	Caged birds (set 2)	13	**0.950**	**0.945**	**0.821**	11.9	**0.942**	**0.54**
	Pianist (set 2)	21	0.912	**0.943**	**0.823**	11.9	**0.950**	**0.33**
	Skier (set 2)	19	0.907	**0.942**	**0.826**	12.9	**0.945**	**0.42**
	Train (set 2)	21	0.924	**0.927**	**0.794**	11.8	**0.951**	**0.48**
Sets of references	Set 1	73	0.940	**0.948**	0.800	10.8	0.954	0.47
	Set 2	90	0.893	0.895	0.714	16.1	0.909	**0.64**

[a]Each row corresponds to a subset of n images either spanning a particular range of quality or corresponding to a particular distortion. The Pearson linear correlation r, the Spearman rank correlation ρ, and the Kendall rank correlation τ are computed between the perceived quality and perceived utility scores. The RMSE and the OR were computed using the utility scores and the mapped [i.e., Eq. (1)] quality scores. r_{fit} denotes the Pearson linear correlation after applying the mapping. For the correlation statistics and OR, bold values are statistically equivalent to the largest value for a subset of images (excluding All). Bold RMSE values are statistically larger than the other subsets based on ANOVA.

after applying the nonlinearity. For each statistic, values in boldface are statistically greater than those of the other levels within that factor. The following summarizes key observations, which appear in bold, followed by statistical justifications and interpretations.

Quality does not consistently and accurately predict utility for different regions of quality. The entire collection of distorted images range from unrecognizable to visually lossless, and a strong global correlation is observed, which implies that a poor-quality image is less useful than an excellent-quality image. However, the 95% PI for the fitted nonlinear mapping between utility and quality (i.e., Fig. 4) indicates that a perceived quality score corresponds to a broad range of perceived utility scores, and the range of the perceived utility scores varies for different regions of quality (e.g., the PI is wider in the medium-quality region than the low-quality region). An analysis of the relationship between the perceived utility scores and the perceived quality scores for individual quality regions provides more insight into the relationship between quality and utility.

For different quality regions, both the correlation and accuracy between the perceived utility scores and the nonlinearly mapped perceived quality scores vary. The perceived utility scores and perceived quality scores exhibit the most linear relationship ($r = 0.82$) for images with low quality (i.e., rated as having either "bad" or "poor" perceived quality). Variations in perceived quality scores explain 67% (i.e., $100r^2\%$) of the variation in perceived utility scores in this quality region. However, for the other quality regions, the correlation between perceived utility scores and perceived quality scores is statistically significantly smaller ($r < 0.62$), which indicates that variations in the perceived quality scores explain no more than 40% of the variation in the perceived utility scores in the medium- and high-quality regions.

The quality region was found to be a factor that influences the squared errors between the perceived utility scores and the nonlinearly mapped perceived quality scores based on a one-way ANOVA ($F(2, 160) = 7.16$, $p < 0.01$). The MSE between the perceived utility scores and the mapped perceived quality scores for distorted images in the medium-quality region is statistically larger than that of the other two quality regions ($p \leq 0.01$).

The significant variation in both the correlation and accuracy statistics for different regions of quality demonstrate that quality does not generally provide a reliable estimate of utility. The observed relationship between quality and utility is discussed for each quality region.

Variations in quality for distorted images in the low-quality region largely coincide with variations in utility. The slope of the overall relationship between utility and quality decreases as quality increases and is steepest within the low-quality region, which indicates that small changes in perceived quality in the low-quality region affect perceived utility more than small changes in quality for other regions of quality. Consider, for example, a reference/distortion sequence beginning with an unrecognizable image and evolving toward a useful image with medium perceived quality. Subsequent images in the sequence will contain less distortion than the previous images, and the sequence will evolve from unrecognizable to recognizable within the low-quality region. The strong correlation ($r = 0.82$) as well as the steep slope between utility and quality within this region reflect the dramatic perceptual changes coinciding with the evolution of images from unrecognizable to recognizable in this sequence. In other words, the observed relationship between quality and utility in the low-quality region suggests that observers largely judge lower-quality images in terms of their ability to interpret the content.

Distorted images in the medium-quality region are useful, but visibly distorted and nearly span the full range of utility: $[21, 115]$. Of the distorted images in the medium-quality region, 20% have very high utility (i.e., perceived utility scores greater than 90) and span nearly the entire range of the medium-quality region: $[2.5, 3.7]$. This clearly demonstrates that high utility does not necessarily imply high quality, since these images all have medium quality. Therefore, very useful images can contain a moderate amount of visible distortions (i.e., have medium quality). Further analysis revealed that most of the images with medium quality and high utility are TS + HPF distorted images, which suggests that removing low-frequency content can form a perceptually different image (i.e., decrease quality) without affecting the image's usefulness.

Distorted images in the high-quality region contain few visible distortions and span a narrow range of utility: $[73, 108]$. In addition, more than 60% of the distorted images have very high utility (i.e., perceived utility scores greater than 90) with quality as low as 4 (i.e., "good" quality). Furthermore, both low correlation with and low RMSE between the perceived utility scores and the nonlinear mapped perceived quality scores was observed for distorted images in the high-quality region. In other words, as the level of distortion decreases utility saturates before quality saturates, and refinements in quality for high-quality images have little effect on utility.

The interpretation of the relationship between utility and quality must be qualified with respect to the natural images used in the experiments. In particular, the usefulness of the natural images was determined by an object or objects that generally occupy a large portion of the image, which led to useful images despite the presence of visible of distortions (i.e., images in the medium-quality region). Had the usefulness of the images been dictated by either a smaller or less conspicuous object (e.g., recognition of the flower pot in the boy and cat image), the relationship between utility and quality could differ. For example, image usefulness dictated by a smaller, inconspicuous object is expected to require a higher quality image than if the usefulness is dictated by a larger, conspicuous object. Such variations in image usefulness reflect tasks that repurpose the original intent of the images. In this paper, the task was to report the content of each natural image, and the content of the images selected for the experiment is dictated by one or two conspicuous objects.

Utility is not accurately estimated using quality for TS + HPF distorted images. Both the accuracy with which perceived utility scores are estimated from mapped perceived quality scores as well as the correlation between the perceived utility scores and the perceived utility scores varies among the different distortion types [54]. The squared errors between the perceived utility scores and the mapped perceived quality scores were influenced by the distortion type factor based on a one-way ANOVA ($F(4, 158) = 3.43$, $p = 0.01$). The MSEs for estimates of perceived utility scores from perceived quality scores for TS + HPF distortions were

found to be statistically larger than those for JPEG ($p < 0.04$), J2K + DCQ ($p < 0.05$), and TS distortions ($p < 0.03$).

TS + HPF distortions disrupt both high-frequency content via TS and low-frequency content via high-pass filtering, whereas JPEG, J2K + DCQ, and TS distortions primarily disrupt high-frequency content before low-frequency content. The perceived utility scores exhibit very strong correlation ($r > 0.93$) with the perceived quality scores for the JPEG, J2K + DCQ, and TS distorted images, and the highest correlation is observed for the TS distorted images ($r = 0.96$). The very strong correlation between the perceived utility scores and the perceived quality scores for JPEG, J2K + DCQ, and TS distorted images indicates that distortions to high-frequency content affect both utility and quality. However, the correlation between the perceived utility scores and the perceived quality scores is statistically lower for the TS + HPF distorted images than the TS distorted images ($p = 0.01$), yet the TS + HPF distorted images only lack the low-frequency content of the TS distorted images. The weak correlation as well as the large RMSE between the perceived utility scores and the mapped perceived quality scores for TS + HPF distorted images indicate that distortions to low-frequency content affect utility differently than they affect quality.

Overall, the analysis of the relationship between utility and quality demonstrate that an image with low quality also has low utility, and an image with high quality also has high utility. However, distorted images with quality in the medium region correspond to a wide range of perceived utility scores, including high utility. In other words, high utility does not imply high quality. The perceived utility scores of TS + HPF distorted images are less accurately estimated from the perceived quality scores than for the other distortions, especially when the TS + HPF distorted image has quality in the medium region and suggests that low-frequency content affects quality differently than utility.

Quality does not accurately predict utility for some reference images. The accuracy with which perceived utility scores are estimated from mapped perceived quality scores varies among the different reference images. As reported in Table 2, the squared errors between the perceived utility scores and the mapped perceived quality scores were not influenced by the reference image based on a one-way ANOVA ($F(8, 154) = 1.68, p = 0.11$). However, when sets of reference images were compared to one another, significant differences in the squared errors between the perceived utility scores and the mapped perceived quality scores were noted ($F(1, 161) = 9.48, \ p < 0.01$). The reference images were grouped into the two sets: Set 1 = {*airplane*, *backhoe*, *guitarist*, *jackolanterns*} and Set 2 = {*boycat*, *cagedbirds*, *pianist*, *skier*, *train*}.

The accuracy with which perceived utility scores were estimated from mapped perceived quality scores was significantly lower for the reference images in set 1 than those in set 2. Specifically, the TS + HPF distorted images generated from reference images in set 2 were generally rated as having perceived quality scores much lower than their TS distorted image counterparts (i.e., equal γ). In other words, observers were more sensitive to the loss of low-frequency content in image from set 2 than for images from set 1.

B. Effects of Low-Frequency Content on Quality and Utility

JPEG, BLOCKS, J2K + DCQ, and TS distortions largely disrupt high-frequency content with limited disruption to low-frequency content. However, TS and TS + HPF distorted images with the same smoothing parameter γ only differ with regard to the inclusion of low-frequency content. The perceived utility scores and perceived quality scores for TS and TS + HPF distorted images were compared to determine the influence of low-frequency content on both utility and quality.

For each reference image, the subjective scores for TS and TS + HPF distorted images with equal smoothing parameters γ are tested for statistical differences when $\gamma = 1$, 5, 21, 97, 446, and 2048. Statistical differences in the subjective scores imply that the disruption to low-frequency content influences the subjective scores. For TS and TS + HPF distorted images formed from the same reference image using smoothing parameter γ, let $S_{\mathrm{TS}(\gamma)}$ and $S_{\mathrm{TS+HPF}(\gamma)}$ denote the subjective scores, respectively, and let $\sigma_{S_{\mathrm{TS}(\gamma)}}$ and $\sigma_{S_{\mathrm{TS+HPF}(\gamma)}}$ respectively denote the standard deviation of $S_{\mathrm{TS}(\gamma)}$ and $S_{\mathrm{TS+HPF}(\gamma)}$. z tests were used to determine if two scores are statistically different using the test statistic

$$z_{\mathrm{stat}} = \frac{S_{\mathrm{TS}(\gamma)} - S_{\mathrm{TS+HPF}(\gamma)}}{\sqrt{\sigma^2_{S_{\mathrm{TS}(\gamma)}} + \sigma^2_{S_{\mathrm{TS+HPF}(\gamma)}}}}. \tag{2}$$

The results of the z test are reported as the confidence that $S_{\mathrm{TS}(\gamma)}$ is greater than $S_{\mathrm{TS+HPF}(\gamma)}$ (i.e., $P(z \le z_{\mathrm{stat}})$, where z is a zero-mean Gaussian random variable with unit variance) and is denoted as $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)}) \in [0, 1]$. Figures 6 and 7 present $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ as a function of the

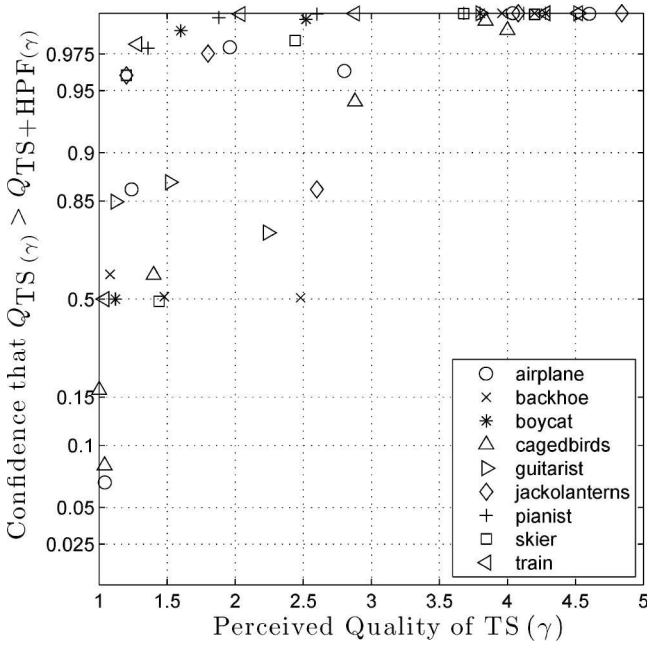


Fig. 6. Perceived quality either decreases or remains the same when low-frequency content is disrupted (i.e., for TS + HPF distortions relative to TS distortions). The figures show the confidence that the perceived quality (Q) score of the TS distortions are greater than the perceived quality score for TS + HPF distortions with equal γ as a function of the perceived quality score of the TS distortions. See Subsection 4.B for additional details regarding the confidence analysis and its interpretation.

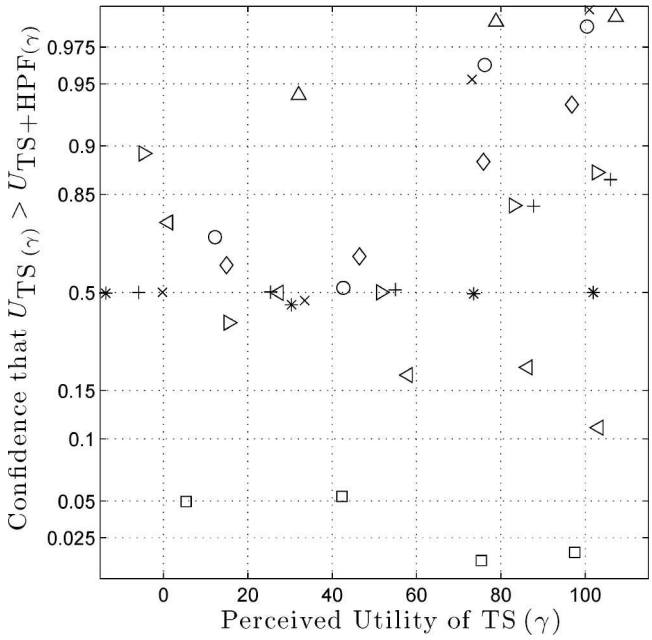


Fig. 7. Disruptions to low-frequency content do not affect the perceived utility of most images. The figures show the confidence that the perceived utility (U) score of the TS distortions are greater than the perceived utility score for TS + HPF distortions with equal γ as a function of the perceived utility score of the TS distortions. Refer to the caption of Fig. 6.

perceived quality score and the perceived utility score of a TS distorted image, respectively [55]. Key observations appear in bold, followed by a statistical justification and interpretation.

For the same reference image, a TS + HPF distorted image never is of higher quality than a TS distorted image with the same γ. Over all levels of quality, loss of low-frequency content led to an average decrease in perceived quality of 0.53, and, in most cases, the perceived quality of a TS distorted image is statistically greater than that of a TS + HPF distorted image formed from the same reference image using the same γ. For some images, the perceived quality of a TS and TS + HPF distorted image with equal γ are statistically equivalent but only when the perceived quality of the TS distorted image is less than 3 (i.e., the quality is "fair" or worse). In short, because poorer-quality images are very heavily distorted, additional distortions that affect the low-frequency content of poorer-quality images have little influence on the perceived quality.

The relationship between the utility of TS and TS + HPF distorted images with the same γ formed from the same reference image varies for each reference image. For many of the reference images, disruptions to low-frequency content (i.e., TS and TS + HPF distorted images with equal γ) do not affect perceived utility. However, disruptions to the low-frequency content of the skier, airplane, backhoe, and caged birds images did affect utility when the TS distorted image has high utility (i.e., perceived utility score greater than 70).

The skier image has a statistically greater perceived utility score when low-frequency content is disrupted (i.e., for TS + HPF distorted images) than when the low-frequency content is not disrupted (i.e., the TS distorted images). Moreover, a skier TS + HPF distorted image with medium quality has a perceived utility score statistically greater than 100: this image is more useful than the reference image. Removing the low-frequency content from the skier image introduces "halos" near edges that enhance the visibility of the skier and other objects (see Fig. 8). The increased visibility of the skier could explain why removing the low-frequency content (i.e., a TS distorted image versus a TS + HPF distorted image with the same γ) increased the perceived utility. However, the observer

(a) TS ($\gamma = 1$), $Q = 4.2$, $U = 98$

(b) TS+HPF ($\gamma = 1$), $Q = 3.4$, $U = 115$

Fig. 8. Example showing that the skier TS distorted image has statistically greater quality than the TS + HPF distorted image with equal γ but statistically lower utility. Removing the low-frequency content from the skier image (i.e., the TS + HPF distorted image) introduces "halos" near edges that enhance the visibility of the skier. See also Figs. 6 and 7.

responses do not indicate what criteria the observers used to choose the TS + HPF distorted image over the TS distorted image (see Subsection 7.A).

Among TS distorted images with high utility (i.e., greater than 70), the perceived utility scores of the airplane, backhoe, and caged birds images were statistically smaller for TS + HPF distorted images than TS distorted images for the same γ. Because a paired comparison test methodology without ties was used, observers were forced to choose one of the images in each pair presented. The binary responses collected from observers to obtain perceived utility scores preclude a definitive explanation for why the TS distorted images were chosen over TS + HPF distorted images, but there are two possible explanations for this result:

- Relative to the TS + HPF distorted images, the low-frequency content of TS distorted images may convey useful information about the content to observers. For example, in the airplane image, the removal of the low-frequency content darkens many regions of the image (e.g., the sky and the airplane). The sky similarly darkens in the backhoe image when low-frequency content is removed. These perceptual differences may cue different interpretations about the scene to observers, and the interpretation for the TS distorted image appears more accurate. The appearance of the specular reflections of the bird cage, which may provide an observer with information about the brightness of the room, are reduced in the caged birds TS + HPF image relative to its TS distorted version. Such features correspond to additional information about the image content beyond the visibility of the objects' spatial details, which would be primarily conveyed by high-frequency content (e.g., edges).
- Observers may have found both TS and TS + HPF distorted images formed from the same reference using the same γ equally useful and more often reverted to judgments of quality to choose an image. This would suggest that quality is a secondary criteria to utility. In other words, given images with equal utility, observers generally preferred the higher-quality TS distorted image, except when the lower-quality TS + HPF distorted image conveyed sufficiently more information about the content (e.g., the skier image). For many of the reference images, the values of $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ show evidence of a slight, though not statistically significant, bias toward observers choosing the TS distorted image over the TS + HPF distorted image with equal γ.

We conjecture that the second explanation (i.e., observers revert to quality judgements) is more plausible; however, different observers may have used different criteria to make a decision (see Subsection 7.A).

C. Examples Illustrating That Quality Is Not a Proxy for Utility

The analysis of the relationship between perceived utility scores and perceived quality scores demonstrates that quality does not accurately predict utility, and Fig. 9 illustrates several cases when the relationship between two distorted images based on quality does not reflect the relationship between those two images in terms of utility and vice versa. Each row of Fig. 9 corresponds to a different reference image, and for each row the images are arranged such that (1) the distorted image on the left and the distorted image in the middle have statistically equivalent perceived utility scores but statistically different perceived quality scores and (2) the distorted image in the middle and the distorted image on the right have statistically equivalent perceived quality scores but statistically different perceived utility scores.

The first two rows of the first two columns in Fig. 9 illustrate the relationship between TS and TS + HPF distorted images. The TS parameter γ must be increased (i.e., increasing the level of TS) for a TS distorted image to exhibit the same perceived quality observed as a TS + HPF distorted image, but the resulting TS distorted image will have lower perceived utility than the TS + HPF distorted image (first row of Fig. 9). Similarly, a J2K + DCQ distorted image that exhibits the same perceived quality as a TS + HPF distorted image also has lower perceived utility (second row of Fig. 9). In other words, high-frequency content must be disrupted to form a distorted image with equal quality to an image that lacks low-frequency content.

The last row of Fig. 9 contains three images that respectively have J2K + DCQ, JPEG, and TS + HPF distortions. High-frequency content is disrupted for both J2K + DCQ and JPEG distorted images with limited disruption to low-frequency content. For the TS + HPF distorted image, the low-frequency content is lost with little disruption to the high-frequency content. The TS + HPF distorted image has "fair" perceived quality (statistically equivalent to the JPEG distorted image) but perceived utility corresponding to the REC.

These examples illustrate that distorted images corresponding to a specific level of utility can significantly vary in terms of quality, and distorted images corresponding to a specific level of quality can significantly vary in terms of utility. Thus, quality does not reliably predict utility. Furthermore, the observed relationship between utility and quality implies that any objective estimator that accurately estimates perceived quality (utility) scores cannot also accurately estimate perceived utility (quality) scores across a variety of distortion types.

5. OBJECTIVE ESTIMATORS OF SUBJECTIVE SCORES

This section reviews several signal analysis tools that could provide meaningful estimates of subjective scores of natural images: (1) amplitude–spectrum statistics of natural images, (2) natural image quality estimators, and (3) a proposed natural image utility estimator that compares image contours.

A. Amplitude–Spectrum Statistics

A well-known characteristic of natural scenes is the relationship between the spatial frequency and the amplitude of the spatial frequency component [56]. This characteristic is mathematically specified as $A(f) = f^{-\beta}$, where β defines the spectral slope of an image. Natural images have been reported to have spectral slope values near 1.2 on average [56,57].

Human performance on visual discrimination tasks has demonstrated a decrease when the spectral slope of the test stimuli are artificially increased or decreased [57]. Such results motivate the use of the spectral slope as an indicator of perceived utility as a natural image is increasingly distorted. In this paper, the spectral slope β of a test image is evaluated as a means to estimate subjective scores.

Fig. 9. Differences in perceived quality (Q) do not imply differences in perceived utility (U). In terms of perceived utility, the distorted images in the middle column are statistically equivalent to the distorted images in the left column. However, in terms of perceived quality the distorted images in middle column are statistically equivalent to the distorted images in the right column. The images have been cropped from their original versions.

B. Full-Reference Image Quality Estimators

The psychometric evidence presented in Section 4 establishes that perceived quality scores do not reliably predict perceived utility scores. This evidence implies that an objective estimator that accurately estimates perceived quality scores cannot accurately estimate perceived utility scores. Accurate estimation of the perceived quality of distorted natural images remains an open research problem, so current quality estimators may produce accurate estimates of the perceived utility scores of distorted natural images. Therefore, full-reference quality estimators are treated as mathematical formulas and, in particular, signal analysis tools that quantify the comparison of a distorted image to a reference image. This section reviews the full-reference quality estimators assessed according to their performance as utility estimators and quality estimators in Section 6.

Full-reference quality estimators use both an explicit, external reference image $\mathbf{X}$ and the test image $\hat{\mathbf{X}}$ to estimate the subjective score of the test image. The full-reference quality estimators evaluated in this paper can be categorized as (1) conventional signal fidelity measures, (2) estimators based on properties of the HVS, and (3) estimators derived from hypothetical high-level HVS objectives.

1. Conventional Signal Fidelity Measures

MSE, which is used to compute the peak signal-to-noise ratio (PSNR), and rms distortion contrast provide computationally simple evaluations of signal fidelity. These measures evaluate fidelity solely in terms of the overall energy of the distortions. rms distortion contrast $C_{\mathrm{rms}}(\mathbf{E})$ measures fidelity based on the visibility of the distortions $\mathbf{E} = \hat{\mathbf{X}} - \mathbf{X}$ when comparing the images on a particular display device [58] and is given by

$$C_{\mathrm{rms}}(\mathbf{E}) = \frac{1}{\mu_{L(\mathbf{X})}}\left[\frac{1}{M}\sum_{i=1}^{M}(L(E_i+\mu_{\mathbf{X}}) - \mu_{L(\mathbf{E}+\mu_{\mathbf{X}})})^2\right]^{1/2}, \quad (3)$$

where $\mu_{L(\mathbf{X})}$ denotes the average luminance of the reference image $\mathbf{X}$, $L(E_i+\mu_{\mathbf{X}})$ denotes the luminance of the ith pixel of $\mathbf{E}+\mu_{\mathbf{X}}$, $\mu_{L(\mathbf{E}+\mu_{\mathbf{X}})}$ denotes the average luminance of the mean shifted distortions $\mathbf{E}+\mu_{\mathbf{X}}$, and M is the total number of pixels. Equation (3) normalizes the standard deviation of the luminance values $\mathbf{E}+\mu_X$ according to the mean luminance of $\mathbf{X}$. This normalization accounts for Weber's law, which asserts that distortions of equal energy are more difficult to detect in brighter regions of an image than in darker image regions. Various other signal fidelity measures have been analyzed with regard to their performance to estimate perceived quality [59,60].

2. Estimators Based on Properties of the HVS

Several quality estimators capitalize on models and principles characterizing low-level HVS properties such as contrast sensitivity [61], contrast masking [32,61,62], and perceived contrast [63,64]. These properties model the detection of a visual target (e.g., the distortions in an image) under a variety of conditions based on the contrast of the distortions. Many quality estimators have been proposed [15–17,21–23,65–75], but this section summarizes a subset that represents a variety of approaches.

Two quality estimators, the weighted SNR (WSNR) and noise quality measure (NQM), evaluate images by incorporating HVS properties to simulate the appearance of the reference and test images to a human and compute the SNR as a function of the difference of the simulated images [72]. Another quality estimator, the visual SNR (VSNR), evaluates images according to a contrast model accounting for low-level HVS properties and the midlevel HVS property of global precedence [74,76]. The last quality estimator in this category, criterion 4 (C4), assesses images using elaborate models of several processing areas of the visual cortex [74]. The models in C4 describe color vision, frequency-orientation analysis, contour detection, perceptual and localization of patterns, object discrimination, and visual memory.

3. Estimators Based on Hypothetized Objectives of the HVS

A family of quality estimators has been developed based on the premise that the HVS has evolved in response to the statistical regularities exhibited by the physical world. The estimators operate under the hypothesis that differences between the statistical characteristics of the reference and test images correspond to a change in perceived quality. Estimators from this family include the structural similarity (SSIM) index [22], a multiscale extension of SSIM (MS-SSIM) [73], and the visual information fidelity (VIF) criterion [23].

SSIM employs a local measure of spatial correlation between the pixels of the reference and test images that is modulated by distortions quantified by locally normalized first (mean) and second (variance) moments. MS-SSIM extends SSIM by evaluating this modified spatial correlation measure across several image scales. The authors of this paper have reported extended discussions and analyses of SSIM and MS-SSIM elsewhere [40,77].

The VIF criterion [23] generates objective scores based on a measurement of the mutual information between the test and reference image. VIF uses Gaussian models of spatially local wavelet coefficients of the test image and reference image, so the mutual information measurement reduces to a local SNR in the wavelet domain [see Eq. (A3)]. A modification of VIF, denoted VIF*, is also evaluated [78]. VIF* normalizes the individual image scale measurements used by VIF before linearly pooling. Consequently, VIF* exhibits a greater sensitivity to low-frequency content disruptions than VIF. A mathematical description of VIF* is provided in the Appendix A.

C. NICE Utility Estimator

Processing in the HVS parses a visual stimulus into meaningful pieces that facilitate the perception of objects. The primary visual cortex extracts local, oriented edge information from a visual stimulus. This information is later processed by cortical regions of the HVS that have been associated with object perception [79]. Cells within in the extrastriate cortex, in particular V4, have been functionally described as shape descriptors [28]. The extrastriate visual cortex has been shown to exhibit an increased activation in response to images that contain contour information [30]. Thus, the evidence suggests that the HVS uses contour information for object perception.

A degradation to image contours is hypothesized to inhibit object perception. Furthermore, we hypothesize that the perceived usefulness of a distorted image is related to a human's ability to recognize objects within that image. Biderman and Ju reported that human observers can recognize objects from line drawings nearly as efficiently as photographs [80], and the authors of the present paper have shown elsewhere that humans can recognize image content from contour information detected using a Canny edge detector operating at different image scales [81]. The fidelity of contour information from a test image with respect to a reference image may be a reliable indicator of perceived utility, and, specifically in this paper, a human's ability to extract information from the test image.

The NICE utility estimator compares the contours identified in a test image to those identified in the reference image to produce a numerical score indicating the estimated utility score of the test image [78,82]. Image contours or edges, defined by sudden intensity changes in pixel values, can be identified by the presence of an absolute maximum magnitude in the gradient of an image [83].

Image contours can be detected from a single image scale or across multiple image scales. For example, the Sobel edge detector analyzes image content from a single image scale to identify contours. However, energy from edges span multiple image scales, and the HVS does not strictly analyze one image scale of visual information [61]. A wavelet decomposition coarsely approximates the multiscale, multiorientation analysis conducted by the primary visual cortex, and can be used to identify contours at multiple image scales. The Sobel edge detector is computationally efficient, but multiscale contour identification uses visual information from multiple image scales that would be available to the HVS. The performance of NICE was evaluated using both single- and multiscale contour identification methods. The computation that NICE conducts using identified contours is described and followed

by individual descriptions of the single-scale and multiscale contour identification methods used for NICE.

1. Contour Comparison

An objective score with NICE is computed by comparing the contours of the reference and test images, which are represented as binary images. Before the contours of the reference and test images are compared, binary images representing the contour maps are individually subjected to morphological dilation with a 3×3 plus sign-shaped structuring element E [84]. Morphological dilation accommodates local registration errors between the reference and test contour maps introduced by distortions in the test image that should not be quantified as errors.

The contours of the reference and test images are compared across S image scales, and B_s and $\hat{B}_s$ respectively denote the contours of the reference and test images at scale s. The overall NICE score for the test image is

$$\text{NICE} = \frac{\sum_{s=1}^{S} d_H(B_s \bigoplus E, \hat{B}_s \bigoplus E)}{\sum_{s=1}^{S} N_{B_s}}, \tag{4}$$

where N_{B_s} is the number of nonzero elements of $B_s \bigoplus E$, $d_H(X, Y)$ denotes the Hamming distance [85] between the two binary vectors X and Y, and $B \bigoplus E$ denotes the dilation of the binary image B using the morphological structuring element E. The Hamming distance quantifies (1) the number of pixels corresponding to contours in the reference image that have been lost in the test image due to the distortions and (2) the number of pixels corresponding to contours in the test image introduced by the distortions that were absent in the reference image. Since the content of natural images vary, the proportion of pixels corresponding to contours will vary. The factor N_B accounts for this variability by adaptively scaling the raw score $d_H(B \bigoplus E, \hat{B} \bigoplus E)$ according to the extent of the contour information identified in the reference image.

2. Single-Scale Contour Identification with Classical Edge Detectors

Numerous image processing tools have been designed to detect edges in natural images [83,86,87]. These are used to generate the binary images B_1 and $\hat{B}_1$ corresponding to contours of the finest image scale of the respective reference and test images for the single-scale implementation of NICE [i.e., $S = 1$ in Eq. (4)]. Edge detectors incorporate a filtering operation that approximates the first derivative of the image. The Sobel and Canny edge detectors were used for the single-scale version of NICE.

The Sobel edge detector filters an image with two 3×3 linear filters, one that approximates a horizontally oriented derivative and another that approximates a vertically oriented derivative. If G_x and G_y correspond to the approximated horizontal and vertical derivatives of the original image, respectively, then an edge-intensity image, given as $G = G_x^2 + G_y^2$, is subjected to hard thresholding, using a threshold given as twice the average value of G to produce a binary image identifying image contours.

The Canny edge detector filters the image with the derivative of a Gaussian specified for a particular $\sigma > 0$ and applies thresholding to generate a binary image [86]. The parameter σ in the Canny filter controls the suppression of high-frequency content (i.e., textures and uncorrelated noise) before detecting edges, and NICE was implemented with the Canny edge detector for $\sigma = 1$.

3. Multiscale Contour Identification

A wavelet representation of an image provides multiscale directional derivatives of that image, which can be used to identify image contours at different image scales. Both the reference and test images are represented using an undecimated implementation of the steerable pyramid [88] using D orientations and S scales [89]. Let $W_{s,\theta}(i)$ and $\hat{W}_{s,\theta}(i)$ denote the ith wavelet coefficient of the respective reference and test images in the subband corresponding to scale $s \in \{1, 2, \ldots, S\}$ and orientation $\theta \in \{0, \frac{\pi}{D}, \frac{2\pi}{D}, \ldots, \frac{\pi(D-1)}{D}\}$.

For each image scale s, the local modulus maxima (LMM) [90] of wavelet coefficient scales correspond to image contours for the reference and test images. The LMM are determined from gradient vectors formed from wavelet subbands corresponding to derivatives in horizontal and vertical spatial directions [90]. Define $G_s(i) = W_{s,0}(i) - jW_{s,\frac{\pi}{2}}(i)$ and $\hat{G}_s(i) = \hat{W}_{s,0}(i) - j\hat{W}_{s,\frac{\pi}{2}}(i)$ as the gradient of the respective reference and test images at scale s, where $j = \sqrt{-1}$. For image scale s, let $M_s(i) = |G_s(i)|$ and $A_s(i) = \angle G_s(i)$ denote the respective modulus and angle of the gradient of the reference image. Similarly, define $\hat{M}_s(i) = |\hat{G}_s(i)|$ and $\hat{A}_s(i) = \angle \hat{G}_s(i)$ for the test image. The LMM of the reference image correspond to points of $M_s(i)$ greater than the two adjacent neighbors in the direction indicated by $A_s(i)$, and for the test image, the LMM are similarly identified using $\hat{M}_s(i)$ and $\hat{A}_s(i)$. For scale s, let $\mathcal{I}_s$ and $\hat{\mathcal{I}}_s$ denote sets of indices i corresponding to LMM of the respective reference image and test images.

Binary images represent image contours of the reference and test images. Thresholds used to identify contours are independently calculated for the reference and test images based on the energy of the combined horizontal and vertical subbands (i.e., M_s and $\hat{M}_s$). Specifically, the image contours at scale s of the reference and test images are identified as LMM that exceed the respective thresholds $\beta_s = \frac{4}{P}\sum_{i=1}^{P} M_s^2(i)$ and $\hat{\beta}_s = \frac{4}{P}\sum_{i=1}^{P} \hat{M}_s^2(i)$, where P is the number of wavelet coefficients. $B_s(i)$ and $\hat{B}_s(i)$, the reference and test binary images for scale s, are defined as

$$B_s(i) = \begin{cases} 1 & M_s(i) > \beta_s \quad \text{and} \quad i \in \mathcal{I}_s \\ 0 & \text{else} \end{cases}. \tag{5}$$

$\hat{B}_s(i)$ is similarly defined using $\hat{M}_s$, $\hat{\mathcal{I}}_s$, and $\hat{\beta}_s$.

6. RESULTS: OBJECTIVE ESTIMATES OF UTILITY AND QUALITY

Subjective experiments are reliable but prohibitively expensive methods to estimate either utility or quality, but an objective estimator that is consistent with subjective responses for either utility or quality can be used in lieu of the subjective experiments. This section evaluates each objective estimator described in Section 5 as both a utility estimator and a quality estimator. Specifically, the objective estimates are evaluated using the perceived utility and perceived quality scores from the subjective experiments. Objective estimators that provide accurate and reliable estimates of the subjective scores also serve as signal analysis tools that can be analyzed to understand which image characteristics impact the subjective

scores. For example, an objective estimator that reliably estimates perceived utility scores can be dismantled to understand the image characteristics that affect utility.

The implementations of all the objective estimators were obtained from the respective authors and are available in the Metrix Mux compilation of objective estimators [91]. Single-scale implementations of NICE are evaluated using the Sobel and Canny edge detector, respectively denoted as $\text{NICE}_{\text{Sobel}}$ and $\text{NICE}_{\text{Canny}}$. Multiscale implementations of NICE are evaluated using up to four scales [i.e., for $S = 1, 2, 3, 4$ in Eq. (4)], where each implementation is denoted MS-NICE_S (i.e., MS-NICE_3 denotes MS-NICE using the first three image scales).

A monotonic, nonlinear mapping between objective estimates and subjective scores is often recommended before analyzing the performance of an objective estimator [92]. However, the nonlinear mapping functionally compensates for objective estimator's shortcomings and obscures the relationship between the image characteristics analyzed by that objective estimator and those that affect the subjective scores. Thus, a linear mapping between the objective estimates and the subjective scores was used to avoid drawing erroneous conclusions from the results that are due to the nonlinear mapping and not the objective estimator. Furthermore, objective estimators that estimate either utility or quality using only a linear mapping are preferred, since training data is not needed to calibrate the nonlinear mapping associated with the objective estimator (see also Appendix VI.3 of [93]).

An affine linear function $h_{\mathcal{E}}$ that maps the objective estimates to the range of values corresponding to the subjective scores that lie in the domain $\mathcal{E}$ was fitted to the data. The parameters of $h_{\mathcal{E}}$ were found by minimizing the sum of the set of squared residuals $\{(h_{\mathcal{E}}(d_i) - e_i)^2\}_{i=1}^n$ for the n images, where d_i and e_i respectively denote an objective estimate and a subjective score for image i.

To test the performance of an objective estimator as a utility estimator and a quality estimator both correlation and accuracy statistics were used to quantify the relationship between its objective estimates and the respective subjective scores. Specifically, (1) the objective estimates and the subjective scores must be strongly correlated and (2) the objective estimator must accurately estimate the subjective scores.

The correlation and accuracy statistics used in Subsection 4.A (i.e., ρ, τ, r, RMSE, and OR) are used to evaluate the ability of the objective estimators to estimate subjective scores. The resolving power ($\text{RP}_{0.05}$) is another accuracy statistic that is used to specify the smallest difference in fitted objective scores for a pair of test images such that the difference is significant based on the estimated error of the subjective scores at the 95% confidence level [94].

The skewness and kurtosis of the set of residuals $\{h_{\mathcal{E}}(d_i) - e_i\}_{i=1}^n$ are also reported. Values of skewness and kurtosis that differ from 0 and 3, respectively, suggest that the residuals do not come from a Gaussian distribution. The best performing objective estimators will have residuals that come from a Gaussian distribution with a small standard deviation (i.e., small RMSE); such estimators analyze important image characteristics that describe the variation in the subjective scores.

Statistical differences in accuracy are determined by comparing the variance of the residuals corresponding to different objective estimators. An F test frequently is used to compare the variance of the residuals corresponding to different objective estimators, but an assumption with the F test is that the residuals come from a Gaussian distribution [53,92]. For most objective estimators, the residuals did not come from a Gaussian distribution according to the JB normality test [51], so the Brown–Forsythe–Levene (BFL) test [95], rather than the F test, was used to compare the variance of the residuals for different objective estimators, with results reported by the corresponding p value. With the BFL test, p values greater than 0.05 indicate that the variance of the residuals for two estimators are statistically equivalent at the 95% confidence level.

The results that characterize the performance of the objective estimator as both (1) utility estimators and (2) quality estimators are reported separately. A general summary of the results is presented.

A. Results: Objective Estimates of Perceived Utility

A utility estimator should both detect recognizable images and provide accurate estimates of perceived utility.

1. Determining If Test Images Are Recognizable

Objective estimators can be used to determine if test images are recognizable by applying an appropriate threshold to the score generated by that estimator.

An image is either recognizable or unrecognizable. Cast as a two-class detection problem, the performance of an estimator as a detector can be characterized by its receiver operating characteristic (ROC) [96–98]. A ROC curve summarizes the relationship between the proportion of true positives and false-positives for a given estimator using a range of threshold values. The area under the ROC curve (AUC) collapses the performance of an objective estimator to a single number. Given a pair of test images belonging to each class (i.e., one recognizable and one unrecognizable), the AUC quantifies the probability that an estimator correctly distinguishes recognizable images from unrecognizable images.

The objective estimators were evaluated as recognition detectors by applying a threshold to the objective estimates to classify an image as either recognizable or unrecognizable. A total of 1000 thresholds were tested ranging from 0.95 of the minimum objective estimate to 1.05 times the maximum objective estimate. For each threshold, the true positive rate (i.e., the proportion of times an image was correctly classified as recognizable) and the false-positive rate (i.e., the proportion of times an image was incorrectly classified as recognizable) were recorded. ROC curves were generated from the recorded pairs of true-positive and false-positive rates. The AUC was estimated by the trapezoidal rule [97]. The AUC is a statistic estimated from available data and is therefore a random variable, so the 95% confidence intervals for the estimates of the AUC were computed [97]. The first column of Table 3 lists the AUC as the recognition detection accuracy for each objective estimator that was used to detect recognizable images across all distortions.

VIF, VIF*, $\text{NICE}_{\text{Sobel}}$, $\text{NICE}_{\text{Canny}}$, and all versions of MS-NICE correctly distinguish recognizable images from unrecognizable images with statistically greater probability than the other objective estimators. All of the other objective estimators correctly rank two such images with probability greater than chance. In Table 3, the absolute maximum value of the recognition detection accuracy is shown in bold, and values

Table 3. Statistics Summarizing the Performance of Estimators as Utility Estimators[a]

			Estimating Perceived Utility							
			Correlation Measures			Accuracy Measures				
	Estimator	Recognition Detection Accuracy	ρ	τ	r	RMSE	OR	$RP_{0.05}$	BFL_p	Skew/Kurt
Spectral slope	β	0.729	0.751	0.535	0.730	25.6	0.748	64.4	$<10^{-3}$	0.51/2.8
Signal fidelity measures	PSNR	0.768	0.520	0.422	0.414	34.1	0.859	57.3	$<10^{-3}$	−0.19/2.6
	$C_{rms}(\mathbf{E})$	0.792	0.521	0.404	0.211	36.6	0.877	38.2	$<10^{-3}$	0.11/1.8
Estimators based on HVS properties	WSNR	0.766	0.485	0.372	0.415	34.0	0.847	57.6	$<10^{-3}$	−0.22/2.4
	NQM	0.796	0.509	0.401	0.422	33.9	0.847	54.1	$<10^{-3}$	−0.28/2.4
	VSNR	0.790	0.530	0.436	0.541	31.5	0.742	83.9	$<10^{-3}$	−0.51/3.0
	C4	0.830	0.661	0.517	0.651	28.4	0.785	75.9	$<10^{-3}$	−0.74/3.9
Estimators based on hypothesized HVS objectives	SSIM	0.924	0.862	0.682	0.845	20.0	*0.595*	55.2	$<10^{-3}$	−0.12/3.8
	MS-SSIM	0.935	0.731	0.585	0.652	28.4	0.828	66.4	$<10^{-3}$	0.01/2.4
	VIF	*0.978*	**0.959**	**0.821**	**0.943**	**12.4**	0.595	**26.6**	**1**	0.04/2.9
	VIF*	*0.973*	0.928	*0.768*	*0.924*	14.3	*0.497*	41.1	*0.850*	−0.53/4.2
Proposed utility estimators	$NICE_{Sobel}$	*0.980*	*0.951*	*0.804*	*0.924*	14.3	0.564	33.6	*0.398*	−0.37/4.1
	$NICE_{Canny}$	*0.980*	*0.937*	*0.785*	*0.935*	13.3	**0.454**	39.1	*0.472*	−0.36/5.2
	MS-$NICE_1$	*0.979*	*0.956*	*0.816*	*0.923*	14.4	0.583	33.0	*0.296*	−0.35/3.7
	MS-$NICE_2$	*0.980*	*0.959*	*0.821*	*0.911*	15.4	0.577	33.4	*0.073*	−0.15/3.6
	MS-$NICE_3$	*0.980*	*0.958*	*0.817*	0.902	16.2	0.601	34.0	0.016	−0.06/3.5
	MS-$NICE_4$	**0.981**	*0.947*	*0.794*	0.901	16.3	0.601	34.5	0.008	0.03/3.3

[a]The recognition detection accuracy is the probability that an unrecognizable image and a recognizable image are correctly distinguished. The Pearson (linear) correlation coefficient r, the Spearman rank correlation coefficient ρ, the Kendall rank correlation τ, the RMSE, the OR, and the resolving power $RP_{0.05}$ are reported when the estimates are compared with the perceived utility scores for test images with perceived utility exceeding −15 ($n = 163$ test images). Italicized p values for the BFL test (BFL_p) indicate that the residual variance is statistically equivalent to that of VIF. The skewness and kurtosis of the residuals are italicized when the JB test indicates that the residuals belong to a Gaussian distribution (see Section 6). Except for the skewness and kurtosis statistics, optimal values appear in bold with statistically equivalent values italicized.

that are statistically equivalent with 95% confidence are italicized. The subjective experiments revealed a linear relationship between perceived quality scores and perceived utility scores for low-quality distorted images, so an objective estimator that produces accurate estimates of perceived quality scores should also accurately detect recognizable images. All the other objective estimators exhibit poor recognition detection accuracy because these estimators severely underestimate the perceived utility scores of TS + HPF distorted images. Specific details about the performance of these estimators are discussed alongside the results presented in Subsection 6.A.2.

2. Estimating the Perceived Utility of Recognizable Test Images

A utility estimator should accurately estimate the perceived utility of a test image deemed recognizable. Only those test images with perceived utility scores exceeding −15 ($n = 163$ test images) are used to evaluate an estimator's performance as a utility estimator, since accurate estimates of perceived utility scores for unrecognizable images are unnecessary. Table 3 summarizes the correlation and accuracy statistics for all the objective estimators when analyzing their linearly mapped objective estimates with respect to the perceived utility scores. The p value for the BFL test BFL_p is reported when the residuals of each objective estimator were compared with the residuals of VIF, since residuals for VIF exhibited the smallest variance when VIF was evaluated as a utility estimator.

The following reports the key results, which appear in bold, followed by a summary of the results for subsets of objective estimators that exhibit similar performance. Statistical justifications, general interpretations, and specific remarks about the objective estimators are reported.

Estimators that strictly analyze distortions to high-frequency content and measure degradations to image contours accurately estimate perceived utility. VIF, $NICE_{Sobel}$, $NICE_{Canny}$, and MS-$NICE_{S\leq 2}$ [99] outperform the other objective estimators as utility estimators. Relative to the other estimators evaluated, estimates from these estimators strongly correlate with the perceived utility scores ($r > 0.91$, $\rho > 0.93$, $\tau > 0.78$). Estimates from these objective estimators more accurately estimate the perceived utility scores than the other estimators (RMSE ≤ 15.4, OR < 0.6, $RP_{0.05} < 39.2$).

VIF, $NICE_{Sobel}$, $NICE_{Canny}$, and MS-$NICE_{S\leq 2}$ strictly analyze the high-frequency content of the reference and test images. $NICE_{Sobel}$, $NICE_{Canny}$, and MS-$NICE_{S\leq 2}$ primarily analyze disruptions to contours, whereas VIF analyzes any disruption to high-frequency content (i.e., both contours and textures). Most importantly, all of these estimators do not analyze disruptions to low-frequency content, which contributed to the poorer performance of many of the other objective estimators as utility estimators. A detailed discussion that compares VIF to NICE is presented in Subsection 7.B.

Among the various implementations of NICE and MS-NICE, estimates from $NICE_{Canny}$ most accurately estimate the perceived utility scores. The RMSE for $NICE_{Canny}$ is smallest among the various implementations of NICE and MS-NICE, but is not statistically significant. However, the residuals for $NICE_{Canny}$ exhibit much higher kurtosis that those for the other implementations of NICE and MS-NICE. Residuals exhibiting high kurtosis indicate that most of the estimates from $NICE_{Canny}$ are very accurate with respect to the

perceived utility scores and poorly estimated for only a few distorted images. Further inspection of the relationship between estimates from $\text{NICE}_{\text{Canny}}$ and the perceived utility scores revealed that $\text{NICE}_{\text{Canny}}$ less accurately estimates the perceived utility scores for distorted images formed from the skier and caged birds images relative to distorted images formed from the remaining seven images. Removing distorted images formed from the skier and caged birds images, both significantly increases the linear correlation and significantly reduces the RMSE to 0.97 and 9.3, respectively. The interpretation of none of the other estimators changes as significantly when these distorted images are removed; even the RMSE for VIF only reduces to 11.

$\text{NICE}_{\text{Canny}}$ underestimates the perceived utility scores for the skier distorted images. The Canny edge detector identifies contours within the snow region below the skier in the *skier* image. Because all of the distortions blur the pixel values in the snow region of the image, $\text{NICE}_{\text{Canny}}$ no longer detects most of these contours in the snow region in any of the distorted images at the lowest level of distortion. Consequently, $\text{NICE}_{\text{Canny}}$ measures a large degradation to image contours in these slightly distorted images. Furthermore, a majority of the contours detected in the reference image correspond to the snow region of the image, so additional degradations to contours have a small impact on the estimate from $\text{NICE}_{\text{Canny}}$. The Sobel edge detector did not identify any contours in the snow region of the image, and thus removing skier distorted images from the data set did not change the interpretation of its performance as a utility estimator.

$\text{NICE}_{\text{Canny}}$ overestimates the perceived utility scores for the caged birds distorted images. The cage in the caged birds image blocks the two birds, and the bars of the cage contribute strong edges that are identified by the Canny edge detector. As this image is distorted, the strong edges corresponding to the bars of the cage are not significantly suppressed, and thus, $\text{NICE}_{\text{Canny}}$ only measures a small overall degradation to the image contours. Because the cage partially occludes the birds, a higher-level, more complex analysis is necessary to distinguish the birds from the cage and measure the degradation of their respective contours. We hypothesize that the human observers primarily attend to the birds with an awareness of the cage, and perceived utility is gauged by the detail of the birds. $\text{NICE}_{\text{Canny}}$ does not separately measure the degradation of contours corresponding to the birds and the cage within this image.

For the remaining distorted images, $\text{NICE}_{\text{Canny}}$ outperforms the other implementations of NICE and MS-NICE, and these different implementations largely vary with respect to the edge-detector used. The Sobel, Canny, and wavelet-based edge detectors used by NICE were evaluated using the publicly available Berkeley Segmentation Dataset and Benchmark to determine which method identifies contours that best corresponds with those identified by humans [100]. The wavelet-based edge detector was tested using only its finest scale contour maps (i.e., $s = 1$), since MS-NICE_1 exhibits the smallest residual variance among the four versions of MS-NICE. The Canny edge detector ranked highest among the three methods, which suggests that its contour maps correspond best with those formed by humans. NICE is designed assuming that degradation to contours coincide with a decrease in utility, and better correspondence between the objectively identified contours and those identified by a human should improve the performance of NICE. The overall performance of $\text{NICE}_{\text{Canny}}$ as a utility estimator combined with the correspondence between its contour maps and those identified humans illustrate the importance of contour information when estimating perceived utility.

A monotonic, nonlinear mapping improves the accuracy of MS-NICE_3 and MS-NICE_4 as utility estimators. Estimates from both MS-NICE_3 and MS-NICE_4, strongly correlate with perceived utility scores ($r \approx 0.9$, $0.95 < \rho < 0.96$, $0.79 < \tau < 0.82$), and their rank correlation statistics are statistically equivalent to those of VIF. However, these two estimators produce less accurate estimates of perceived utility (RMSE ≈ 16). A monotonic, nonlinear mapping, which does not affect ρ and τ, improved both the linear correlation and accuracy between estimates from both MS-NICE_3 and MS-NICE_4 and the perceived utility scores. This nonlinear mapping primarily compresses differences among the objective estimates for distorted images with low perceived utility scores (i.e., near the RT). Although the nonlinearity improves their performance as utility estimators, the nonlinear mapping introduces a stage of processing that was not incorporated into MS-NICE_S and illustrates that MS-NICE_S's analysis of the reference and test images for $S > 2$ without the monotonic, nonlinearity degenerates as utility decreases. In particular, MS-NICE_S becomes increasingly sensitive to disruptions to low-frequency content for distorted images with low perceived utility scores as S increases and coarser image scales are analyzed.

VIF* produces unreliable estimates of perceived utility, especially for TS + HPF distortions with high perceived utility. Estimates from VIF* strongly correlate with and accurately estimate perceived utility scores, and most of VIF*'s correlation and accuracy statistics are statistically equivalent to those of VIF. However, VIF* underestimates the perceived utility of TS + HPF distorted images with high perceived utility because, unlike VIF, VIF* has a greater sensitivity to disruptions to low-frequency content. The negative skewness of VIF*'s residuals are a consequence its poor estimates of the perceived utility scores for TS + HPF distorted images. The results from the subjective experiments described in Section 4 demonstrate that disruptions to low-frequency content do not consistently affect perceived utility scores. Therefore, VIF*'s unreliable performance as a utility estimator, especially for TS + HPF distorted images, is expected because VIF* is sensitive to disruptions to low-frequency content.

Estimators that analyze distortions to low-frequency content perform poorly as utility estimators. The spectral slope, signal fidelity measures, objective estimators based on HVS properties, SSIM, and MS-SSIM perform poorly as utility estimators. Estimates from these estimator exhibit weaker correlation with perceived utility scores ($\rho < 0.86$, $\tau < 0.68$, $r < 0.85$) and less accurately estimate perceived utility (RMSE > 20, OR > 0.6, $\text{RP}_{0.05} > 54$) than the other estimators (i.e., the variants of NICE and VIF).

The TS + HPF distorted images largely influence the performance of these estimators. When each estimator was analyzed as a utility estimator with the TS + HPF distorted images removed, all estimators except the spectral slope exhibited significantly better performance as utility estimators. The performance improvements when the TS + HPF distorted images

are removed indicate that these estimators operate with the assumption that distortions do not compromise the integrity of the low-frequency content without also severely distorting the high-frequency content. Such an assumption is consistent with the behavior of lossy image compression methods but could be problematic for other types of distortion artifacts that arbitrarily distort an image such as transmission errors due to packet loss.

The spectral slope quantifies the shape of the distorted image's frequency response. The J2K + DCQ, TS, and TS + HPF distortions primarily disrupt and suppress high-frequency content before low-frequency content as the level of distortion increases, which leads to a significant decrease in the spectral slope (i.e., β increases in $A(f) = 1/f^{-\beta}$). JPEG distortions simultaneously disrupt, suppress, and introduce high-frequency content (e.g., blocking artifacts) and lead to a modest increase in β relative to the other distortions as the level of distortion increases. As a result, the relationship between the spectral slope and perceived utility varies with each distortion type, and the spectral slope is observed to be an unreliable indicator of utility, since its relationship with perceived utility scores varies with distortion type.

The signal fidelity measures as well as the estimators based on HVS properties generate objective estimates that are entirely, or in part, a function of energy measurements of the reference and test images. PSNR and $C_{\mathrm{rms}}(\mathbf{E})$ measure the global energy of the difference image $\mathbf{X} - \hat{\mathbf{X}}$ in the pixel and luminance domains, respectively. VSNR analyzes the visibility of the global contrast of the difference image across several image scales. The other estimators based on HVS properties apply different filters to suppress frequency content less sensitive to the HVS and compare the global energy of the filtered reference and test images in the frequency domain. All of these estimators account for distortions to low-frequency content, and the loss of low-frequency content significantly decreases the energy of the distorted image relative to the reference image. Consequently, each of these estimators underestimate the perceived utility scores for TS + HPF distorted images.

Both SSIM and MS-SSIM incorporate an analysis of low-frequency content via a comparison of the spatially local mean pixel values of the reference and test images. In addition to MS-SSIM's local mean comparison of the reference and test images, MS-SSIM compares the variance of spatially local pixel values of the reference and test images across multiple image scales. Thus, both MS-SSIM's mean and variance comparisons analyze the low-frequency content of the reference and test images, whereas only SSIM's mean comparison analyzes the low-frequency content of the reference and test images.

SSIM and MS-SSIM were modified by removing the comparisons of the reference and test images that quantify disruptions to low-frequency content, and both modified estimators exhibited better performance as utility estimators than their original implementations across all five distortion types. The linear correlation and RMSE between SSIM's estimates and perceived utility significantly improve to 0.92 and 15, respectively, when SSIM operates without the local mean comparison (i.e., when SSIM ignores disruptions to low-frequency content). The linear correlation and RMSE between MS-SSIM's estimates and perceived utility modestly improve to 0.73 and 25, respectively, when MS-SSIM operates without both the local mean and variance comparisons across multiple image scales. Even when the local mean and variance comparisons have been removed, MS-SSIM's multiscale analysis necessarily quantifies distortions to low-frequency content and explains its modest performance improvement. However, the significant improvement demonstrated with SSIM when the local mean comparisons are removed relative to the original implementation of SSIM suggests that an analysis of high-frequency content provides reliable estimates of perceived utility [101].

B. Results: Objective Estimates of Perceived Quality

A quality estimator should produce objective estimates that are both strongly correlated with perceived quality and accurately estimate perceived quality. All test images ($n = 243$) were used to evaluate an estimator's performance as a quality estimator because a reliable quality estimator should accurately determine the quality of unrecognizable distorted images, even though they have "bad" quality. Table 4 summarizes the statistics for each objective estimator when analyzing the linearly mapped objective estimates with respect to the perceived quality scores. The difference between VIF*'s estimates and the perceived quality scores exhibited the smallest variance (i.e., smallest RMSE), so the p value for the BFL test is reported when the residuals of estimates from each objective estimator when used as quality estimates were compared with that of VIF*.

The following reports the key results, which appear in bold, followed by a summary of the results for subsets of objective estimators that exhibit similar performance. Statistical justifications, general interpretations, and specific remarks about the objective estimators are reported.

Estimators that are sensitive to distortions to low-frequency content perform poorly as quality estimators over a variety of distortions. The spectral slope, signal fidelity measures, and objective estimators based on HVS properties, SSIM, and MS-SSIM perform poorly as quality estimators over a variety of distortions. Estimates from these estimators, weakly correlate ($\rho \in [0.52, 0.87]$, $\tau \in [0.33, 0.70]$, $r \in [0.40, 0.88]$) with and/or inaccurately estimate (RMSE $\in [0.50, 1.1]$, OR $\in [0.51, 0.89]$, and $\mathrm{RP}_{0.05} \in [1.6, 2.6]$) the perceived quality scores. A difference of 1 in perceived quality corresponds to a different quality category (i.e., "fair" versus "good").

The TS + HPF distortions are largely responsible for the poor performance of these estimators as quality estimators. In fact, when each estimator was analyzed with the TS + HPF distortions removed from the test image set, the interpretation of the performance of these estimators changes: the correlation and accuracy statistics of these estimators improved. Apart from the spectral slope and $C_{\mathrm{rms}}(\mathbf{E})$, these objective estimators previously have been evaluated as quality estimators on other image databases that do not include distortions that deliberately disrupt the low-frequency content without severely disrupting the high-frequency content [74,102,103]. The performance of these estimators on the current database of test images, which includes distortions that disrupt low-frequency content without severely disrupting high-frequency content (i.e., the TS + HPF distortions for small γ), demonstrates that these estimators were designed

Table 4. Statistics Summarizing the Performance of Objective Estimators as Quality Estimators[a]

		Correlation Measures			Accuracy Measures				
	Estimator	ρ	τ	r	RMSE	OR	$RP_{0.05}$	BFL_p	Skew/Kurt
Spectral slope	β	0.518	0.331	0.585	0.895	0.835	1.902	$<10^{-3}$	–0.27/2.1
Signal fidelity measures	PSNR	0.598	0.477	0.656	0.833	0.506	1.949	$<10^{-3}$	–0.81/2.8
	$C_{rms}(\mathbf{E})$	0.627	0.480	0.401	1.011	0.881	2.413	$<10^{-3}$	–0.61/2.0
Estimators based on HVS properties	WSNR	0.582	0.443	0.648	0.841	0.823	2.052	$<10^{-3}$	–0.90/2.8
	NQM	0.600	0.461	0.666	0.823	0.831	1.911	$<10^{-3}$	–0.97/3.0
	VSNR	0.607	0.466	0.738	0.745	0.794	1.760	$<10^{-3}$	–1.1/3.6
	C4	0.822	0.636	0.832	0.615	0.808	1.600	$<10^{-3}$	–0.47/2.9
Estimators based on hypothesized HVS objectives	SSIM	0.870	0.696	0.883	0.519	0.700	2.517	$<10^{-3}$	–0.12/2.6
	MS-SSIM	0.713	0.561	0.603	0.850	0.864	1.918	$<10^{-3}$	–0.38/1.9
	VIF	*0.929*	*0.774*	*0.950*	0.345	**0.531**	**0.828**	*0.13*	0.17/5.4
	VIF*	*0.938*	*0.799*	**0.959**	**0.313**	*0.568*	1.056	**1**	0.12/3.0
Proposed utility estimators	$NICE_{Sobel}$	*0.932*	*0.780*	0.885	0.515	0.786	2.076	$<10^{-3}$	–0.64/2.9
	$NICE_{Canny}$	*0.914*	*0.746*	0.934	0.394	0.568	1.020	*0.35*	–0.29/3.5
	MS-$NICE_1$	*0.935*	*0.784*	0.875	0.535	0.778	2.256	$<10^{-3}$	–0.77/3.1
	MS-$NICE_2$	*0.937*	*0.789*	0.860	0.563	0.765	2.405	$<10^{-3}$	–0.79/3.1
	MS-$NICE_3$	*0.940*	*0.796*	0.855	0.572	0.782	2.291	$<10^{-3}$	–0.73/3.0
	MS-$NICE_4$	**0.946**	**0.810**	0.855	0.572	0.757	2.254	$<10^{-3}$	–0.69/3.0

[a]The Pearson (linear) correlation coefficient r, the Spearman rank correlation coefficient ρ, the Kendall rank correlation τ, the RMSE, the OR, and the resolving power $RP_{0.05}$ are reported when the estimates are compared with the perceived quality scores for all test images ($n = 243$). Italicized p values corresponding to the BFL test (BFL_p) indicate that the residual variance is statistically equivalent to that of VIF*. The skewness and kurtosis of the residuals are italicized when the JB test indicated that the residuals belong to a Gaussian distribution (see Section 6). Except for the skewness and kurtosis statistics, optimal values appear in bold with statistically equivalent values italicized.

and tested under the assumption that either (1) distortions will not compromise the integrity of the low-frequency content, (2) distortions to low-frequency content will coincide with severe distortions to high-frequency content, or (3) distortions to low-frequency content have a negligible impact on quality. However, the current results indicate that these different assumptions do not reflect the general image characteristics that influence judgments of perceived quality. Namely, the loss of low-frequency content without severely disrupting high-frequency content coincides with a significant decrease in quality.

The spectral slope, as discussed in Subsection 6.A.2, quantifies the shape of the distorted image's frequency response, which varies for the different distortions. However, the correlation between the spectral slope and the perceived quality scores is significantly lower than the correlation between the spectral slope and the perceived utility scores. Specifically, the spectral slope accounts for 53% (i.e., $100r^2\%$) of the variation of utility, but only 34% of the variation in quality. An analysis of the relationship between the spectral slope and the perceived quality scores revealed that TS + HPF distorted images have spectral slopes similar to TS and J2K + DCQ distorted images, but TS + HPF distorted images have significantly lower perceived quality. Thus, the spectral slope is an unreliable indicator of quality over a variety of distortions.

The signal fidelity measures as well as the estimators based on HVS properties, excluding C4, produce estimates that are a function of the energy of the reference and test images and account for distortions to low-frequency content, which, according to the subjective experiments, significantly affects quality. However, these estimators are very sensitive to distortions to low-frequency content and consequently underestimate the perceived quality scores of TS + HPF distorted images.

An analysis of the relationship between the estimates from C4, SSIM, and MS-SSIM and the perceived quality scores revealed that their accuracy decreases as quality decreases, which indicates that their analyses of the reference and test images degenerate as quality decreases. However, the Spearman rank correlation ($\rho > 0.70$) between perceived quality and the estimates from these three estimators suggest that they each exhibit a nonlinear, monotonic relationship with the perceived quality scores. Fitting the estimates from these estimators to the perceived quality scores with a monotonic, nonlinear mapping significantly changes the interpretation of their performance as quality estimators: each significantly improves as a quality estimator. Each of these estimators analyze distortions to low-frequency content, as discussed in Subsection 6.A.2, and the subjective experiments demonstrate that distortions to low-frequency content affect perceived quality. However, even with a nonlinear mapping these estimators remain sensitive to distortions to low-frequency content and still underestimate the perceived quality of TS + HPF distorted images.

Estimators that analyze all frequency content without overemphasizing the significance of distortions to low-frequency content accurately estimate perceived quality scores over a variety of distortions. VIF* produces more reliable estimates of perceived quality scores than VIF over a variety of distortions. Estimates from VIF strongly correlate ($\rho > 0.92$, $\tau > 0.77$, $r > 0.95$) with and accurately estimate (RMSE < 0.35, OR < 0.57, $RP_{0.05} \in [0.83, 1.1]$) perceived quality scores, and most of VIF's correlation and accuracy statistics are statistically equivalent to those of VIF*.

VIF distinguishes smaller differences among distorted images with high perceived quality more reliably than VIF*, which results in smaller resolving powers for VIF because VIF is more sensitive to disruptions to high-frequency content than VIF*. Modest disruptions to high-frequency content (i.e.,

textures) affect the perceived quality of high-quality yet visibly distorted images. However, distortions to low-frequency content have a greater affect on perceived quality than distortions to high-frequency components (see Section 4), and VIF* is more sensitive to low-frequency distortions than VIF. Consequently, VIF* estimates the perceived quality scores of TS + HPF distortions more accurately than VIF, which results in the slightly smaller, although not statistically significant, RMSE observed for VIF* as compared to VIF. However, VIF overestimates the perceived quality scores of TS + HPF distorted images because disruptions to low-frequency content do not affect estimates from VIF unless they accompany severe disruptions to high-frequency content. VIF*, however, analyzes the low-frequency content. In short, VIF performs well as a quality estimator for applications that do not encounter distortions such as the TS + HPF distortions that disrupt low-frequency content without severely disrupting high-frequency content. However, VIF* performs well as a quality estimator across a variety of distortions because its modifications to VIF normalize the individual channel measurements based on the energy distribution of the reference image across image scales (see Section 8).

Estimators that measure degradations to image contours perform poorly as quality estimators over a variety of distortions. $\text{NICE}_{\text{Sobel}}$ and the various implementations of MS-NICE produce unreliable estimates of perceived quality across a variety of distortions. Estimates from these estimators strongly correlate ($\rho \in [0.91, 0.95]$, $\tau \in [0.74, 0.81]$, $r \in [0.85, 0.94]$) with and estimate with moderate accuracy ($\text{RMSE} \in [0.39, 0.58]$, $\text{OR} \in [0.56, 0.79]$, $\text{RP}_{0.05} \in [1.0, 2.5]$) the perceived quality scores.

A nonlinear relationship between the perceived quality scores and the estimates from both $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ was observed and quantified by their strong Spearman correlation statistics ($\rho > 0.93$). Further analysis of this nonlinear relationship revealed that small degradations to contours, as measured by both $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$, correspond to large changes in the perceived quality scores. In other words, distorted images with high perceived quality scores primarily exhibit visible degradations to textures, and both $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ do not measure degradations to image textures, which influence perceived quality. Furthermore, distorted images with very low perceived quality exhibit large changes in contours, as measured by $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$, but exhibit very little change in perceived quality. Thus, heavily distorted images (i.e., with very low perceived quality) exhibit strong variations in signal characteristics that correspond to very small changes in perceived quality. This follows if one considers again a reference/distortion sequence beginning with an unrecognizable image and evolving toward a useful, medium-quality image. The dramatic perceptual changes in subsequent images near the RT will coincide with significant variations in the underlying signal characteristics, especially the emergence of contours, as detected by $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$. Despite these dramatic perceptual changes, the perceived quality scores of these images are still very low relative to the undistorted reference images.

For $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$, a monotonic, nonlinear mapping increases the linear correlation between their objective estimates and the perceived quality scores to at least 0.94 and is statistically larger for MS-NICE_4 ($r = 0.97$). The nonlinear mapping also reduces the RMSE to less than 0.41 and is smallest for MS-NICE_4 ($\text{RMSE} = 0.28$). The fitted nonlinearity expands small differences among estimates from $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ for distorted images with high perceived quality and compresses large differences among estimates from $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ for distorted images with low perceived quality. Among the single- and multiscale implementations of NICE, MS-NICE_4 exhibits the best performance as quality estimator when fitted with a nonlinear mapping because, as discussed in Subsection 6.A.2, implementations of MS-NICE_S for larger S are more sensitive to low-frequency distortions than the other versions (i.e., NICE and MS-$\text{NICE}_{S\leq 2}$), which analyze distortions to high-frequency content.

Although the monotonic, nonlinear mapping changes the interpretation of the performance of $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ as quality estimators, the parameters of this nonlinearity may vary for distortions not included in the current collection of test images. The current results cannot definitively establish that using both $\text{NICE}_{\text{Sobel}}$ and MS-$\text{NICE}_{S\leq 4}$ with a tuned nonlinear mapping provides reliable and accurate estimates of perceived quality over a variety of distortion types.

$\text{NICE}_{\text{Canny}}$ performs poorly as a quality estimator for medium-quality distorted images. Over the entire collection of distorted images, estimates from $\text{NICE}_{\text{Canny}}$ exhibit correlation and accuracy statistics as a quality estimator that are statistically equivalent to those of VIF* when considering the entire collection of distorted images. However, the performance of $\text{NICE}_{\text{Canny}}$ as a quality estimator is not consistent for different regions of quality. Specifically, estimates from $\text{NICE}_{\text{Canny}}$ exhibit statistically weaker linear correlation with the perceived quality scores ($r = 0.62$) than VIF* ($r = 0.82$) for distorted images with medium quality (i.e., perceived quality scores between $[2.25, 3.75]$). Furthermore, the RMSE between estimates using both VIF* and $\text{NICE}_{\text{Canny}}$ and perceived quality scores are 0.28 and 0.42, respectively, for medium-quality distorted images, and the variance of the residuals are statistically smaller for VIF* than $\text{NICE}_{\text{Canny}}$. In both the low- and high-quality regions, the performance statistics for VIF* and $\text{NICE}_{\text{Canny}}$ are statistically equivalent.

The relationship between $\text{NICE}_{\text{Canny}}$ and the perceived quality scores is consistent with the relationship observed between perceived quality scores and perceived utility scores: perceived utility is unreliably predicted from perceived quality for medium-quality distorted images. Likewise, $\text{NICE}_{\text{Canny}}$ estimates the perceived quality less reliably for distorted images with medium quality. TS + HPF and TS distorted images with equal γ formed from the same reference image have very similar values for $\text{NICE}_{\text{Canny}}$, which is consistent with their equal perceived utility scores yet different perceived quality scores. $\text{NICE}_{\text{Canny}}$ overestimates the quality of TS + HPF distorted images because it does not analyze distortions to low-frequency content, whereas VIF* does and most accurately estimates the perceived quality of TS + HPF distorted images.

C. Results: Summary

When estimating perceived utility scores, objective estimators that analyze the high-frequency content of the reference and test images outperform those estimators that also analyze the low-frequency content of the reference and test images. Specifically, VIF, $\text{NICE}_{\text{Sobel}}$, $\text{NICE}_{\text{Canny}}$, and MS-$\text{NICE}_{S\leq 2}$ produce

the most reliable estimates of perceived utility scores. The interpretation of both SSIM and MS-SSIM as utility estimators changes when they operate without the components that analyze low-frequency content (i.e., the mean component and, in the case of MS-SSIM, also the variance component): both estimators provide more accurate estimates of perceived utility than their original implementations.

$NICE_{Canny}$ produces the most accurate estimates of the perceived utility scores when the skier and caged birds images were discarded. These images reveal two limitations of $NICE_{Canny}$: (1) detection of less visible contours (e.g., those in snow region in the skier image) and (2) separate analysis of relevant versus irrelevant contours (e.g., the birds versus the bars of the cage in the caged birds image). Despite these limitations, $NICE_{Canny}$ demonstrates that perceived utility scores can be reliably estimated from an analysis of image contour degradation.

When estimating perceived quality scores, estimates from VIF* most accurately estimate the perceived quality scores. Unlike many of the other objective estimators, VIF* analyzes both high- and low-frequency content of the reference and test images without overemphasizing disruptions to low-frequency content. Several other estimators grossly underestimate the perceived quality scores of TS + HPF distorted images because these estimators analyze low-frequency content but overemphasize the effect of distortions to low-frequency content. VIF* weights the relative influence of distortions to low- and high-frequency content on its estimates in a manner that yields accurate estimates of perceived quality.

7. DISCUSSION

The subjective experiments establish that perceived quality is not a suitable proxy for perceived utility. An evaluation of objective estimators as both utility and quality estimators revealed that an analysis of degradations to high-frequency content and, specifically, image contours produces accurate estimates of perceived utility, whereas a properly weighted analysis of degradations across all frequency content produces accurate estimates of perceived quality. This section discusses (1) the limitation of the perceived utility scores, (2) the image characteristics revealed by objective estimators that impact perceived utility and perceived quality, and (3) the relationship between object recognition, perceived utility, and the analysis conducted by NICE [104].

A. Limitations of Perceived Utility Scores

Relative perceived utility scores of distorted images were obtained using a paired comparison methodology that has two limitations. The subjective responses lack information about the specific content actually recognized by the observers viewing the distorted images because the test method only collected binary responses (i.e., a choice) from observers in response to the query, "Which image tells you more about the content?" This precludes an analysis of the data based on the actual criteria that led observers to their responses.

The second limitation is that observers may have used a secondary factor such as perceived quality to choose an image when both images appeared equal with regard to their perceived usefulness. For example, for the airplane, backhoe, and caged birds images, the TS distorted images had higher perceived utility than the TS + HPF distorted image with the same γ. If observers consistently rely on a secondary factor to choose an image, then the perceived utility scores will be intermixed with these secondary factors. Because TS distorted images have greater perceived quality than TS + HPF distorted images, the perceived quality is the most likely secondary factor to influence an observer's decision.

Despite the limitations with the current method used to obtain relative perceived utility scores, the results still illustrate a distinction between perceived quality and perceived utility, and any improvements to the test methodology used to obtain relative perceived utility scores are expected to reveal greater differences between perceived quality and perceived utility.

B. Objective Estimators Reveal Image Characteristics That Impact Utility and Quality

Among the objective estimators investigated, VIF and NICE performed best as utility estimators, and VIF* performed best as a quality estimator. First, the signal analyses conducted by VIF* and VIF are analyzed and compared, since the distinctions between VIF* and VIF reiterate the conclusion drawn from the subjective experiments that low-frequency content affect perceived utility but not quality. Second, the signal analyses conducted by VIF and NICE are analyzed and compared, since VIF and NICE illustrate different uses of high-frequency content to estimate utility. Last, the impact that an edge detector used with NICE has on its performance as a utility estimator for other distortions is discussed.

1. VIF Versus VIF: Low-Frequency Content Affects Quality*

VIF and VIF* analyze the reference and test images using the steerable pyramid decomposition [88], which models the well-accepted multichannel characterization of the analysis conducted by the HVS in the primary visual cortex [61] (a mathematical description of VIF and VIF* is presented in Appendix A). VIF and VIF* compute and linearly pool spatially local SNRs within each channel, which produces a channel measurement that quantifies the fidelity of the test image with respect to the reference image within that channel. The channel measurement values decrease as the fidelity of the test image with respect to the reference image within that channel decreases (i.e., the test image contains more distortion). The sum of the channel measurements from the same image scale yield image scale measurements that quantify the fidelity of the test image with respect to the reference image within that image scale. Because the steerable pyramid decomposition represents a coarser image scale with half as many coefficients as the next finest image scale (i.e., due to decimation), the finer image scale measurements are larger than the coarser image scale measurements. VIF linearly pools image scale measurements to produce an objective estimate for the test image, and image scale measurements at finer image scales dominate VIF's objective estimate. In contrast, VIF* normalizes each image scale measurement by the number of coefficients in that image scale, which balances the measurements from different image scale measurements, before linearly pooling. Natural images exhibit a $1/f^{\alpha}$ power spectra [56], and, consequently, the normalized image scale measurements at coarser image scales dominate VIF*'s objective estimate. As a result, VIF* is more sensitive to disruptions to coarser image scale content than finer image scale content.

Images from the airplane/J2K + DCQ, airplane/TS, and airplane/TS + HPF sequences that have statistically equivalent perceived utility are evaluated using VIF and VIF* to illustrate the differences between VIF and VIF*. The image from the airplane/TS + HPF sequence has the same parameter γ as the image from the airplane/TS sequence and statistically has the smallest perceived quality. Figure 10 shows the image scale measurements from VIF and the normalized image scale measurements from VIF* for these three images. The image scale measurements from VIF are much larger at finer image scales (i.e., high spatial frequencies) than coarser image scales (i.e., low spatial frequencies) and exhibit very little variation among these four distorted images across all image scales. Thus, for these images, VIF's pooled image scale measurements reflect their similarity in perceived utility but not their differences in perceived quality. In contrast, the normalized image scale measurements from VIF* are larger at coarser scales than finer scales and indicate a difference between the airplane/TS + HPF image and the other distorted image at the coarsest image scale. Thus, for these images, VIF*'s pooled image scale measurements reflect their differences in perceived quality and not their similarity in perceived utility.

The analyses conducted by VIF* and VIF are consistent with the subjective experiments. The absence of low-frequency content (i.e., the TS + HPF distorted images versus TS distorted images with the same γ) significantly and consistently affects quality but has less consistent effects on the utility. Since VIF and the various implementations of NICE outperform the other objective estimators as utility estimators, the fidelity of low-frequency content does not strongly influence utility in this study. The low-frequency content represents the shading in grayscale natural images, which forms the appearance of naturalness due to interactions between object surfaces and lighting. Natural images with undisrupted shading are visually consistent with our daily experiences with natural environments. Disruptions to an image's shading decrease its perceived quality, which the objective estimates produced by VIF*, not VIF, accurately reflect due to normalizing image scale measurements before pooling across image scales.

2. Comparing VIF and NICE: Estimates of Image Contour Degradation

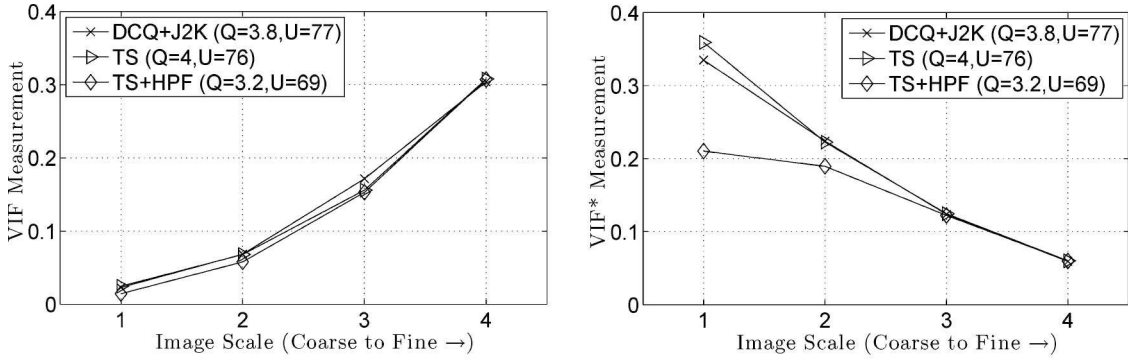
Fine-scale signal components describe natural image details corresponding to both object boundaries and textures, and the energy of the fine-scale signal components coincides with the visibility of these details. VIF and NICE, both of which perform best as utility estimators, specifically analyze the energy of fine-scale signal components of the reference and test images to produce an objective estimate of the test image's perceived utility. Both objective estimators [105] filter the images using two channels that separate the fine-scale signal components into horizontally and vertically oriented spatial frequency components. VIF and NICE illustrate two possible uses of the fine-scale signal components to estimate perceived utility.

VIF subjects the high-frequency channel responses for the reference and test images to a normalization mechanism functionally similar to divisive normalization (i.e., a model of gain control) that normalizes channel responses to a particular range for subsequent processing stages [23,106,107]. Divisive normalization models the relationship between the nth neuron's response y_n to its input t_n according to

$$y_n = \frac{t_n^p}{b^q + \sum_{m \in \mathcal{M}_n} w_m t_m^q}, \tag{6}$$

where b is a positive saturation constant, $\mathcal{M}_n$ is a set of indices specifying local spatial, frequency, and orientation neuron responses to input t_n, the w_m are weights applied to those local responses before pooling, and the exponents p and q are positive values that model a power-law relationship between a neuron's input and output.

VIF approximates the divisive normalization model by normalizing the channel responses based on the energy [i.e., in Eq. (6) set $b = 0$ and $p = q = 2$] of their spatially local channel responses. That is, VIF performs spatially local variance normalization. Image contours generally elicit larger channel responses than textures, and following a spatially local variance normalization, the channel responses to both contours and textures are normalized to the same range. As a



(a) VIF (b) VIF*

Fig. 10. VIF is more sensitive to distortions at finer image scales (i.e., high spatial frequencies) over those at coarser image scales (i.e., low spatial frequencies), whereas VIF* is more sensitive to disruptions to coarser scale content than finer scale content. Figures 10(a) and 10(b) respectively show the image scale measurements computed by VIF and VIF* for the airplane image with J2K + DCQ ($Q = 3.8$, $U = 77$), TS ($Q = 4.0$, $U = 76$), and TS + HPF ($Q = 3.2$, $U = 69$) distortions. These images have statistically equivalent perceived utility, but the perceived quality of the TS + HPF distorted image is statistically smaller than the other two distorted images. The pooled image scale measurements for VIF reflect their similarity in perceived utility but not their differences in perceived quality. The pooled image scale measurements for VIF* reflect their differences in perceived quality not their similarity in perceived utility.

consequence of this normalization, estimates from VIF reflect any disruption to the high-frequency channel responses due to the distortions, so disruptions to both image contours and image textures affect VIF's objective estimates.

In contrast with VIF, NICE detects the edges in the reference and test images and can be viewed as performing spatially global variance normalization, collinear facilitation [29], and hard thresholding. NICE and MS-NICE perform global variance normalization by normalizing the channel responses based on the average channel response energy [108]. Global variance normalization reduces the magnitude of all the channel responses, so channel responses to image contours remain larger than those to textures.

Collinear facilitation describes the perceptual facilitation and suppression of channel responses due to interactions (i.e., connected cells) among spatially local and similarly oriented channel responses and suggests that mechanisms mediate the perception of smooth curves from line segments [109,110]. In particular, studies of human observers report that the detection contrast of a target Gabor patch spatially flanked by two high-contrast Gabor patches is highest (i.e., the target is difficult to detect) when the flanking patches are spatially very close to and have the same orientation as the target, whereas the target detection contrast is lowest (i.e., the target is easy to detect) when the spatial distance between the flanking patches and the target is large and oriented orthogonal to the target patch [109]. Furthermore, the target detection contrast is lowest when the global orientation of the line formed by the three patches coincided with the individual patch orientations [110]. All of the edge detectors used for NICE crudely perform collinear facilitation via a thinning operation that retains local maxima.

Hard thresholding removes low-energy channel responses, which largely coincide with textures, and is hypothesized to represent a decision process performed at a later stage of the HVS corresponding to object perception. Disruptions to image textures have a negligible impact on NICE's objective score, since NICE reflects disruptions to image contours due to the distortion process.

Because NICE primarily measures degradations to image contours, we analyzed estimates of VIF when decomposed into separate fidelity measurements for contours and textures. Specifically, VIF was decomposed as

$$\mathrm{VIF} \approx \mathrm{VIF}_{\mathrm{contour}} + \mathrm{VIF}_{\mathrm{texture}}, \tag{7}$$

where $\mathrm{VIF}_{\mathrm{contour}}$ and $\mathrm{VIF}_{\mathrm{texture}}$ respectively represent VIF evaluated on contour and texture components of an image. Estimates from both $\mathrm{VIF}_{\mathrm{contour}}$ and $\mathrm{VIF}_{\mathrm{texture}}$ were evaluated in terms of their performance as utility estimators. The correlation statistics for $\mathrm{VIF}_{\mathrm{contour}}$ increase relative to those for VIF, whereas all of the correlation statistics for $\mathrm{VIF}_{\mathrm{texture}}$ are statistically smaller than those of VIF. The RMSE of $\mathrm{VIF}_{\mathrm{contour}}$ is 10.7, but the residual variance is statistically equivalent to that of VIF (RMSE = 12.4). However, the RMSE for $\mathrm{VIF}_{\mathrm{texture}}$ is 18.3 and is statistically larger than that of VIF. In short, $\mathrm{VIF}_{\mathrm{contour}}$ accurately estimates the perceived utility scores as a function of the fidelity of the contour information.

In summary, VIF analyzes disruptions to both contours and textures while excluding disruptions to low-frequency content, whereas NICE primarily analyzes disruptions to contours to estimate utility. The performance of $\mathrm{VIF}_{\mathrm{contour}}$ as a utility estimator is parallels the performance of NICE, which corroborates the hypothesis that contour degradations coincide with decreased perceived utility.

3. Edge Detectors Impact the Performance of NICE

NICE operates in conjunction with an edge detector and was assessed using three different edge detectors. As a utility estimator, NICE operating with the Canny edge detector (i.e., $\mathrm{NICE}_{\mathrm{Canny}}$) and excluding the skier and caged birds distorted images outperformed NICE operating with the other edge detectors. The performance of $\mathrm{NICE}_{\mathrm{Canny}}$ as a utility estimator was justified in terms of the agreement of its identified edges with object boundaries identified by humans: compared with human ground truth, the Canny edge detector ranked highest among the three edge detectors (see Subsection 6.A.2). Despite the performance of $\mathrm{NICE}_{\mathrm{Canny}}$ as a utility estimator, the current database does not include distorted artifacts that are uncorrelated with the reference image (e.g., independent, additive white Gaussian noise), and the Canny edge detector frequently identifies false contours as a result of these distortion artifacts.

Correlated distortions influence a human's perception of the distortion level more than uncorrelated distortions (i.e., independent, additive white Gaussian noise) [111,112]. Thus, uncorrelated distortions are expected to have a smaller influence on perceived utility than correlated distortions: human observers can "ignore" moderate levels of uncorrelated distortions. NICE estimates perceived utility as a function of the errors between the reference and test edge maps produced by an edge detector: an edge detected in the reference image but absent in the test image produces an error, and an edge absent in the reference image but detected in the test image produces an error. With NICE, more errors imply lower utility, and perceived utility would be underestimated when the errors are largely due to false contours that humans would "ignore." More advanced edge detectors assess various types of edge cues, including pixel value discontinuities and texture boundaries [113,114], but generally conduct a more complex analysis of an image relative to the edge detectors tested with NICE.

The distortion types used in the experiments were spatially correlated with the reference image, so the current collection of test images cannot be used to evaluate the potential vulnerabilities of the contour detection techniques used by NICE. However, the current results based on correlated distortions demonstrate the feasibility of conducting an image contour comparison to accurately estimate perceived utility. NICE operating with robust edge detectors that do not detect false contours due to uncorrelated noise sources are expected to reliably estimate perceived utility scores for such distortions.

C. Object Recognition, Perceived Utility, and NICE

A perceived utility score quantifies the amount of information a distorted image conveys to a human, where the information of a scene included the objects and activities as well as their respective details. We hypothesize that perceived utility is linked to the level of detail with which objects and activities in the scene are recognized.

Objects in the natural world can be described with varying levels of detail, and object recognition studies using images containing one object have examined the effects of simple image filtering on the level of detail accurately recognized by a

human. Such object recognition studies use the taxonomy of objects proposed by Rosch to distinguish these levels of detail, which Rosch named "levels of abstraction" [115]. As an example, a snare drum can be identified as a musical instrument, a drum, or a snare drum, where Rosch's taxonomy respectively assigns these descriptions to the superordinate, basic, and subordinate levels of abstraction. The object recognition studies demonstrate that humans can reliably recognize an object at the basic level using only low-frequency content, whereas subordinate-level recognition requires more high-frequency content [116,117]. Thus, humans only perceive an object's basic-level details but not its subordinate-level details in a low-pass filtered distorted image, and this result is consistent with low-pass filtering leading to a decrease in perceived utility as subordinate-level object details disappear. The object recognition studies also concluded that humans can reliably recognize an object at both the basic and subordinate levels using only high-frequency content [116,117]. Thus, a high-pass filtered distorted image does not affect the level of detail a human perceives about the object, and this result is consistent with high-pass filtering (i.e., TS versus TS + HPF distorted images with the same γ) often negligibly affecting perceived utility.

Another recent perceptual study of object recognition used natural images containing multiple objects of varying size and demonstrated that the number and accuracy with which humans recognized objects in distorted images decreases as the level of blur increases [118]. Furthermore, the size of the objects accurately recognized decreases as the level of blur increases (i.e., disrupting high-frequency content compromises the recognition of smaller objects). These results are consistent with the criteria proposed by Johnson, which was used to design sensors and display devices [8,119]. The Johnson criteria relates the level of object discrimination to the detectability of a bar pattern of a given spatial frequency. For object recognition, the Johnson criteria states that a human must detect a bar grating with four cycles across the object's minimum dimension [120]. Increasing the number of cycles in the bar grating across the object's minimum dimension allows the object to be more accurately identified. Our perceived utility scores are consistent with this evidence because perceived utility decreases as high-frequency content is removed or distorted.

The object recognition studies demonstrate that loss of high-frequency content but not low-frequency content impairs object recognition performance. This evidence is consistent with our subjective experiments and suggest that our perceived utility scores, rather than perceived quality scores, estimate the amount of information recognized by a human. Such studies and our perceived utility scores provide little guidance toward understanding how information is recognized by a human, and in particular, which underlying image characteristics impact usefulness. However, those objective estimators (i.e., VIF, NICE, and MS-NICE) that accurately estimate perceived utility were dismantled and analyzed to understand those image characteristics that impact usefulness. In particular, NICE and MS-NICE estimate utility based on a measurement of the degradation to image contours in a distorted image with respect to a reference image.

Contours form shapes, and object shape is hypothesized to be a primary cue for object recognition by the HVS [121]. Humans reliably recognize objects from line drawings [80], which provide only object shape cues, and even from degraded line drawings [81,122]. Line drawings abstractly represent object shapes using contours, and humans quickly identify contours formed by Gabor patches aligned along a curved path placed in an image composed of an array of randomly oriented Gabor patches [123]. The ability of humans to recognize objects from abstract contour representations along with their reported ease of detecting contours among clutter support theories of shape-based object recognition.

Another object recognition study collected functional magnetic resonance imaging (fMRI) data for various regions of the visual cortex to understand how the HVS performs object recognition. The fMRI data, which measures variations in blood flow, was collected from both the striate (i.e., primary) and extrastriate cortex when humans viewed images that contained only contour regions, texture regions, or both (i.e., the full image) [30]. In that study, the extrastriate cortex responded greatest when humans viewed images that contain only contour regions. The increased activation due to contour information corroborates theories that object recognition is largely driven by contour information (i.e., shape perception) in natural images.

In summary, NICE performs very well as a utility estimator by extracting, comparing, and quantifying the degradation to image contour information in a distorted image with respect to a reference image. Together, the theories that contour information mediates object recognition and the performance of NICE as a utility estimator demonstrate that NICE is a viable signal analysis tool that estimates the usefulness of distorted natural images.

8. CONCLUSIONS

Natural images from imaging systems supply information that facilitate human observers performing various tasks. This paper examined human performance when performing a broad task with natural images: reporting the content of a distorted image. Novel experiments were conducted to measure the usefulness of distorted natural images in terms of this task. In addition, experiments were conducted to measure the perceived quality of these same distorted natural images. Results from both subjective experiments were compared and revealed the perceived quality does not imply an image's perceived utility. In particular, a distortion that removes low-frequency content from an image demonstrated that perceived utility is largely based on the fidelity of high-frequency content and is less affected by distortions to low-frequency content, whereas distortions to any frequency content affects perceived quality. The observed relationship between utility and quality implies that accurate objective quality (utility) estimators will not accurately estimate perceived utility (quality) for a broad class of distortions.

Several objective estimators, mostly designed to estimate perceived quality with one proposed by the authors to estimate perceived utility, were assessed in terms of their performance as utility and quality estimators. Two estimators were shown to accurately estimate utility. One is the VIF criterion, which is customarily used as a quality estimator. A modification to VIF, denoted VIF*, was proposed that outperforms VIF as a quality estimator on the current database of distorted images. The signal analyses conducted by VIF and VIF* are

consistent with the observations from the subjective experiments. Specifically, VIF primarily analyzes disruptions to high-frequency content and accurately estimates perceived utility but not perceived quality, whereas VIF* exhibits increased sensitivity to low-frequency distortions relative to VIF and analyzes disruptions to all frequency content and accurately estimates perceived quality but not perceived utility.

The NICE utility estimator was also shown to accurately estimate utility. NICE estimates utility as a function of both lost and introduced contour information in a distorted image when compared with a reference image. In contrast with VIF, NICE abstractly represents the reference and test images as contours and compares these contours to estimate utility. NICE was shown to be a viable signal analysis tool to estimate the usefulness of a distorted natural image. This result supports hypotheses about the importance of contour information to the HVS for object perception.

APPENDIX A

The VIF criterion is an extension of the information fidelity criterion (IFC) that incorporates a simple HVS model [23,107]. VIF*, a modified version of VIF, adjusts the relative importance of fidelity measurements computed across spatial frequencies to the overall objective estimate by normalizing VIF's channel measurements before linearly pooling across image scales. VIF* provides accurate estimates of perceived quality for a broader set of distortions than VIF. The calculation of VIF* is specified in terms of IFC and followed by a detailed mathematical description of VIF in terms of IFC.

1. VIF* Specification

VIF extends IFC by modeling the HVS as an additive Gaussian noise source that was conjectured by VIF's authors to model aspects of low-level HVS processing [23]. VIF's assessment of a test image is based on spatially local SNR measurements, computed at multiple image scales, of both the reference and test images contaminated with the modeled, low-level HVS noise. VIF compares wavelet coefficients of the test image to those of reference images.

VIF emphasizes fidelity measurements of finer image scales (i.e., higher spatial frequencies) over those of coarser image scales (i.e., lower spatial frequencies). Thus, VIF is invariant to disruptions to low-frequency content (see Fig. 10), which is functionally due to the variation in the number of coefficients blocks B_k for channels at different image scales. Channels corresponding to finer image scales have more wavelet coefficients than channels corresponding to coarser image scales due to the use of a decimated wavelet transform; for a fixed block size P, the number of coefficient blocks is smaller for channels corresponding to coarser image scales. The proposed modifications of VIF, denoted VIF*, normalizes the channel measurements by the number of blocks B_k for that channel.

Let the elements of the length N_k vector $\mathbf{C}^k$ denotes the wavelet coefficients of the kth channel of the reference image [124]. The elements of the length N_k vectors $\mathbf{E}^k$ and $\mathbf{F}^k$ denote the wavelet coefficients of the kth channel of the respective reference and test images that have been contaminated with visual noise. VIF* is given as

$$\mathrm{VIF}^* = \frac{\sum_{k=1}^{K} \frac{1}{B_k} \mathrm{IFC}(\mathbf{C}^k, \mathbf{F}^k)}{\sum_{k=1}^{K} \frac{1}{B_k} \mathrm{IFC}(\mathbf{C}^k, \mathbf{E}^k)}, \tag{A1}$$

where $\mathrm{IFC}(\mathbf{C}^k, \mathbf{F}^k)$ and $\mathrm{IFC}(\mathbf{C}^k, \mathbf{E}^k)$ are defined as in Eq. (A3). As illustrated in Fig. 10, VIF* produces distinct scores that reflect the changes in the perceived quality scores for these images. In particular, disruptions to low-frequency content affect VIF*'s estimate, whereas VIF's estimate does not. The details of Eq. (A1) are defined in Section 8.

2. VIF Specification

VIF parses each wavelet channel into disjoint blocks composed of P coefficients. The following discussion assumes only one channel, so the superscript k is omitted in the subsequent discussion. Let $\vec{C}_b$ and $\vec{D}_b$ correspond to the bth block of P spatially adjacent coefficients of $\mathbf{C}$ and $\mathbf{D}$, respectively. The bth block of wavelet coefficients in the channel of the reference image may be modeled as a Gaussian scale mixture [125,126] random vector given as $\vec{C}_b = s_b \vec{U}$, where s_b is a positive random scalar and $\vec{U}$ is a zero-mean Gaussian random vector of length P with covariance $\mathbf{K}_{\vec{U}}$. Given s_b, the coefficient block $\vec{C}_b$ is a zero-mean Gaussian random scalar with covariance $s_b^2 \mathbf{K}_{\vec{U}}$, and $\vec{C}_b$ is conditionally independent of $\vec{C}_m$ for all $m \neq b$. VIF relates the bth block of wavelet coefficients of the test and reference images using the linear model $\vec{D}_b = g_b \vec{C}_b + \vec{V}_b$, where g_b is a deterministic scalar defined for each block and $\vec{V}_b$ is a zero-mean Gaussian random vector of length P with covariance matrix $\sigma_{\vec{V}_b}^2 \mathbf{I}$ specified for each block b. Thus, given s_b, the block of coefficients $\vec{D}_b$ is also a Gaussian random vector with covariance $g_b^2 s_b^2 \mathbf{K}_{\vec{U}} + \sigma_{\vec{V}_b} \mathbf{I}$.

Independent zero-mean additive Gaussian noise sources model low-level HVS noise in VIF; coefficients of the reference and test images are contaminated with visual noise. Let $\vec{E}_b$ and $\vec{F}_b$ correspond to the bth block of P spatially adjacent coefficients of $\mathbf{E}$ and $\mathbf{F}$, respectively. The output of the HVS model for the reference image is $\vec{E}_b = \vec{C}_b + \vec{M}_b$, and the output of the HVS model for the test image is $\vec{F}_b = \vec{D}_b + \vec{N}_b$. The terms $\vec{M}_b$ and $\vec{N}_b$ are zero-mean Gaussian random vectors of length P with covariance $\sigma_M^2 \mathbf{I} = \sigma_N^2 \mathbf{I}$, where $\sigma_N^2 = \sigma_M^2$ is the HVS model parameter. Thus, given s_b, the block of coefficients $\vec{E}_b$ is a Gaussian random vector with covariance $s_b^2 \mathbf{K}_{\vec{U}} + \sigma_N^2 \mathbf{I}$, and the block of coefficients $\vec{F}_b$ is also a Gaussian random vector with covariance $g_b^2 s_b^2 \mathbf{K}_{\vec{U}} + \sigma_{\vec{V}_b}^2 \mathbf{I} + \sigma_N^2 \mathbf{I}$.

VIF combines two fidelity measurements to yield an overall assessment of a test image. First, a fidelity measurement comparing the reference coefficients before and after the HVS model value is computed. Second, a fidelity measurement comparing the reference coefficients before the HVS model to the processed coefficients after the HVS model is computed. These two fidelity measurements are computed for each wavelet channel. The ratio of the sum of these fidelity measurements across the channels provides an overall assessment of the test image. Let $\mathbf{s}$ be a length B_k vector whose bth element is s_b. Given $\mathbf{s}$, the VIF value is given by

$$\mathrm{VIF} = \frac{\sum_{k=1}^{K} \mathrm{IFC}(\mathbf{C}^k, \mathbf{F}^k)}{\sum_{k=1}^{K} \mathrm{IFC}(\mathbf{C}^k, \mathbf{E}^k)}. \tag{A2}$$

The terms $\mathrm{IFC}(\mathbf{C}^k, \mathbf{F}^k)$ and $\mathrm{IFC}(\mathbf{C}^k, \mathbf{E}^k)$ are based on IFC [107] and are defined as

$$\mathrm{IFC}(\mathbf{C}^k,\mathbf{F}^k)=\sum_{b=1}^{B_k}\log_2\left(\frac{|g_b^2 s_b^2\mathbf{K}_{\vec{U}}+(\sigma_{\vec{V}_b}^2+\sigma_N^2)\mathbf{I}|}{|(\sigma_{\vec{V}_b}^2+\sigma_N^2)\mathbf{I}|}\right) \tag{A3}$$

and

$$\mathrm{IFC}(\mathbf{C}^k,\mathbf{E}^k)=\sum_{b=1}^{B_k}\log_2\left(\frac{|s_b^2\mathbf{K}_{\vec{U}}+\sigma_N^2\mathbf{I}|}{|\sigma_N^2\mathbf{I}|}\right), \tag{A4}$$

where $|\cdot|$ denotes the matrix determinant and the terms g_b, s_b, $\mathbf{K}_{\vec{U}}$, and $\sigma_{\vec{V}_b}$ vary with k and are computed from $\mathbf{C}^k$ and $\mathbf{D}^k$. For channel k, the term g_b is estimated as the linear regression of block $\vec{D}_b$ on the block $\vec{C}_b$, and the variance of the additive zero-mean Gaussian noise $\vec{V}_b$ is the MSE of the regression.

ACKNOWLEDGMENTS

This work was funded by the National Science Foundation (NSF) under grant CCF-0916471.

REFERENCES AND NOTES

1. In this paper, "natural images" are formed using imaging devices that sense the natural environment over the visible portion of the electromagnetic spectrum (e.g., digital cameras). Computer-generated images and other types of synthetic images are not considered natural images.
2. C. G. Ford, M. A. McFarland, and I. W. Stange, "Subjective video quality assessment methods for recognition tasks," Proc. SPIE **7240**, 72400Z (2009).
3. C. Ford, P. Raush, and K. Davis, eds., *Video Quality in Public Safety Conference* (Institute for Telecommunication Sciences, 2009).
4. A. M. Burton, S. Wilson, M. Cowan, and V. Bruce, "Face recognition in poor-quality video; evidence from security surveillance," Psychol. Sci. **10**, 243–248 (1999).
5. J. K. Petersen, *Understanding Surveillance Technologies* (CRC, 2001).
6. J. P. Davis and T. Valentine, "CCTV on trial: matching video images with the defendant in the dock," Appl. Cogn. Psychol. **23**, 482–505 (2009).
7. J. C. Leachtenauer and R. G. Driggers, *Surveillance and Reconnaissance Imaging Systems* (Artech House, 2001).
8. J. Johnson, "Analysis of image forming systems," in *Image Intensifier Symposium* (Fort Belvoir, 1958).
9. L. M. Biberman, ed., *Perception of Displayed Information* (Plenum, 1973).
10. A. van Meeteren, "Characterization of task performance with viewing instruments," J. Opt. Soc. Am. A **7**, 2016–2023 (1990).
11. J. C. Leachtenauer, "Resolution requirements and the Johnson criteria revisited," Proc. SPIE 1–15 (2003).
12. R. H. Vollmerhausen, E. Jacobs, and R. G. Driggers, "New metric for predicting target acquisition performance," Opt. Eng. **43**, 2806–2818 (2004).
13. J. M. Irvine, B. A. Eckstein, R. A. Hummel, R. J. Peters, and R. Ritzel, "Evaluation of the tactical utility of compressed imagery," Opt. Eng. **41**, 1262–1273 (2002).
14. P. D. O'Shea, E. L. Jacobs, and R. L. Espinola, "Effects of image compression on sensor performance," Opt. Eng. **47**, 013202 (2008).
15. T. Stockham, "Image processing in the context of a visual model," Proc. IEEE **60**, 828–842 (1972).
16. J. L. Mannos, "The effects of a visual fidelity criterion on the encoding of images," IEEE Trans. Inf. Theory **20**, 525–536 (1974).
17. D. Granrath, "The role of human visual models in image processing," Proc. IEEE **69**, 552–561 (1981).
18. H. de Ridder and G. M. Majoor, "Numerical category scaling: an efficient method for assessing digital image coding impairments," Proc. SPIE **1249**, 65–77 (1990).
19. J. A. J. Roufs, "Perceptual image quality: concept and measurement," Philips J. Res. **47**, 35–62 (1992).
20. S. A. Klein, "Image quality and image compression: a psychophysicist's viewpoint," in *Digital Images and Human Vision*, A. B. Watson, ed. (MIT, 1993), pp. 73–88.
21. T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. C. Bovik, ed. (Academic, 2000).
22. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).
23. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process. **15**, 430–444 (2006).
24. The National Imagery Interpretability Rating Scale (NIIRS) has been associated with image quality [7]. However, the NIIRS characterizes an image's quality based on the ability of a photo interpreter to detect, recognize, and identify objects in an image. Various versions of the NIIRS have been designed for specific image applications. The NIIRS is more compatible with the definition of utility used in this paper.
25. H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.
26. D. Chandler, "The CSIQ database," http://vision.okstate.edu/index.php?loc=csiq.
27. A visually lossless image is visually indistinguishable from a reference image.
28. T. M. Murphy and L. H. Finkel, "Shape representation by a network of V4-like cells," Neural Netw. **20**, 851–867 (2007).
29. G. Loffler, "Perception of contours and shapes: low and intermediate stage mechanisms," Vis. Res. **48**, 2106–2127 (2008).
30. S. O. Dumoulin, S. C. Dakin, and R. F. Hess, "Sparsely distributed contours dominate extra-striate responses to complex scenes," NeuroImage **42**, 890–901 (2008).
31. The experiments described in this paper augment the experiments described in previous publications by the authors [40,77,80].
32. D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," J. Opt. Soc. Am. A **20**, 1164–1180 (2003).
33. W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard* (Van Nostrand Reinhold, 1993).
34. "Independent JPEG Group," http://www.ijg.org.
35. International Organization for Standardization, "Information technology—digital compression and coding of continuous-tone still images—requirements and guidelines," ITU-T T.81 (International Telecommunication Union, 1992).
36. D. S. Taubman and M. W. Marcellin *JPEG2000: Image Compression Fundamentals, Standards, and Practice* (Kluwer Academic, 2002).
37. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithm," Physica D (Amsterdam) **60**, 259–268 (1992).
38. G. Steidl, J. Weickert, T. Brox, P. Mrazek, and M. Welk, "On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs," SIAM J. Numer. Anal. **42**, 686–713 (2004).
39. J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," IEEE Trans. Image Process. **14**, 1570–1582 (2005).
40. D. M. Rouse and S. S. Hemami, "Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM," Proc. SPIE **6806**, 680615.1–680615.14 (2008).
41. J. S. Bruner and M. C. Potter, "Interference in visual recognition," Science **144**, 424–425 (1964).
42. R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs I: The method of paired comparisons," Biometrika **39**, 324–345 (1952).
43. D. E. Critchlow and M. A. Fligner, "Paired comparisons, triple comparisons, and ranking experiments as generalized linear models, and their implementation on GLIM," Psychometrika **56**, 517–533 (1991).

44. D. Strohmeier and G. Tech, "Sharp, bright, three-dimensional: open profiling of quality for mobile 3DTV coding methods," Proc. SPIE 75420T (2010).
45. International Telecommunication Union, "Subjective video quality assessment methods for multimedia applications," ITU-U P.910 (International Telecommunication Union, 2008).
46. Numerical category scaling [18], adjective category scale [19], and categorical sort [127] are alternative names describing the ACR test method. The subjective assessment methodology for video quality (SAMVIQ) generally obtains more accurate perceived quality scores and avoids many problems where observers avoid using the ends of the quality scale. Both ACR and SAMVIQ yield very similar perceived quality scores for our collection of distorted images [128].
47. "Multimedia group test plan" (2008), draft version 1.21., http://www.vqeg.org.
48. Prior work in the context of perceived quality often denotes a perceived quality score as a mean opinion score.
49. The perceived quality of unrecognizable images with perceived utility scores less than −15 range from 1 to 1.4 with the average, standard deviation, and median being 1.07, 0.089, and 1.04, respectively.
50. G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. (Iowa State, 1989).
51. C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity, and serial independence of regression residuals," Econ. Lett. **6**, 255–259 (1980).
52. E. C. Fieller, H. O. Hartley, and E. S. Pearson, "Tests for rank correlation coefficients. I," Biometrika **44**, 470–481 (1957).
53. J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 5th ed. (Duxbury, 2000).
54. Only six BLOCK distorted images have perceived utility scores greater than −15, so results corresponding to the BLOCK distorted images provide little insight into the relationship between quality and utility. Furthermore, these images have perceived quality scores in the range $[1, 1.3]$ (i.e., "bad" quality) and perceived utility scores in the range $[-13, 4]$ (i.e., effectively useless).
55. Values of $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ less than 0.025 and greater than 0.975 indicate that the subjective scores for TS and TS + HPF distorted images with equal γ are statistically different at the 95% confidence level (i.e., a two-sided z test). Values of $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ less than 0.05 indicate that the subjective score for the TS distorted image is statistically smaller than the subjective score for a TS + HPF distorted image formed from the same reference image using the same γ at the 95% confidence level (i.e., a one-sided z test). Similarly, values of $\mathrm{Conf}(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ greater than 0.95 indicate that the subjective score for the TS distorted image is statistically greater than the subjective score for a TS + HPF distorted image with the same γ.
56. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," J. Opt. Soc. Am. A **4**, 2379–2394 (1987).
57. C. A. Párraga, T. Troscianko, and D. J. Tolhurst, "The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model," Vis. Res. **45**, 3145–3168 (2005).
58. C. Poynton, "The rehabilitation of gamma," Proc. SPIE **3299**, 232–249 (1998).
59. A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," IEEE Trans. Commun. **43**, 2959–2965 (1995).
60. I. Avcıbaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," J. Electron. Imaging **11**, 206–233 (2002).
61. R. L. De Valois and K. K. De Valois, *Spatial Vision* (Oxford University, 1990).
62. G. Legge and J. Foley, "Contrast masking in human vision," J. Opt. Soc. Am. **70**, 1458–1470 (1980).
63. M. A. Georgeson and G. D. Sullivan, "Contrast constancy: debluring in human vision by spatial frequency channels," J. Physiol. **252**, 627–656 (1975).
64. N. Brady and D. J. Field, "What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns," Vis. Res. **35**, 739–756 (1995).
65. W. A. Pearlman, "A visual system model and a new distortion measure in the context of image processing," J. Opt. Soc. Am. **68**, 374–386 (1978).
66. R. J. Safranek and J. D. Johnston, "A perceptually tuned subband image coder with image dependent quantization and post-quantization data compression," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 1989), pp. 1945–1948.
67. S. J. Daly, "The visible difference predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, ed. (MIT, 1993), pp. 179–206.
68. J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A. B. Watson, ed. (MIT, 1993), pp. 163–178.
69. A. B. Watson, "DCT quantization matrices visually optimized for individual images," Proc. SPIE **1913**, 202–216 (1993).
70. P. Teo and D. Heeger, "Perceptual image distortion," Proc. SPIE **2179**, 127–141 (1994).
71. A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," IEEE Trans. Image Process. **6**, 1164–1175 (1997).
72. N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," IEEE Trans. Image Process. **9**, 636–650 (2000).
73. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems, and Computers* (IEEE, 2003), Vol. 2, pp. 1398–1402.
74. D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," IEEE Trans. Image Process. **16**, 2284–2298 (2007).
75. M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," Signal Process., Image Commun. **23**, 239–256 (2008).
76. D. Navon, "Forest before trees: the precedence of global features in visual perception," Cogn. Psychol. **9**, 353–383 (1977).
77. D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," in *Proceedings of the IEEE International Conference on Image Processing* (IEEE, 2008), pp. 1188–1191.
78. D. Rouse, R. Pepion, S. Hemami, and P. Le Callet, "Image utility assessment and a relationship with image quality assessment," Proc. SPIE **7240** (2009).
79. K. Grill-Spector, "The neural basis of object perception," Curr. Opin. Neurobiol. **13**, 159–166 (2003).
80. I. Biderman and G. Ju, "Surface versus edge-based determinants of visual recognition," Cogn. Psychol. **20**, 38–64 (1988).
81. D. M. Rouse and S. S. Hemami, "Quantifying the use of structure in cognitive tasks," Proc. SPIE **6492**, 64921O (2007).
82. D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Proceedings of the IEEE International Conference on Image Processing* (IEEE, 2009), pp. 2217–2220.
83. W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3rd ed. (Wiley, 2001).
84. C. Giardina and E. Dougherty *Morphological Methods in Image and Signal Processing* (Prentice Hall, 1998).
85. The Hamming distance counts the number of dissimilar elements between two vectors [129].
86. D. Marr and E. Hildreth, "Theory of edge detection," Proc. R. Soc. Lond. Ser. B **207**, 187–217 (1980).
87. J. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-8**, 679–698 (1986).
88. E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation," in *Proceedings of the IEEE International Conference on Image Processing* (IEEE, 1995), Vol. 3, pp. 444–447.
89. The high-pass residual generated by the steerable pyramid is not used.

90. S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," IEEE Trans. Pattern Anal. Mach. Intell. **14**, 710–732 (1992).
91. M. D. Gaubatz, D. M. Rouse, and S. S. Hemami, "MeTriX MuX," http://foulard.ece.cornell.edu/gaubatz/metrix_mux.
92. Video Quality Experts Group, "VQEG final report of FR-TV phase II validation test" (2003), http://www.vqeg.org.
93. Video Quality Experts Group, "Final report from the VQEG on the validation of objective models of multimedia quality assessment, phase I," (2008), version 2.6., http://www.vqeg.org.
94. M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1," Signal Process., Image Commun. **19**, 101–107 (2004).
95. M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," J. Am. Stat. Assoc. **69**, 364–367 (1974).
96. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Peninsula, 1988).
97. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology (Oak Brook, Ill.) **143**, 29–36 (1982).
98. T. Fawcett, "An introduction to ROC analysis," Pattern Recogn. Lett. **27**, 861–874 (2006).
99. The notation MS-NICE$_{S\leq 2}$ is used to refer to both MS-NICE$_1$ and MS-NICE$_2$.
100. D. Martin, C. Fowlkes, D. Tal and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the 8th International Conference of Computer Vision* (IEEE, 2001), pp. 416–423.
101. The local variance comparison used by SSIM corresponds to an analysis of high-frequency content and does not need to be removed.
102. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process. **15**, 3440–3451 (2006).
103. E. C. Larson and D. M. Chandler, "The most apparent distortion: a dual strategy for full reference image quality," Proc. SPIE **7242**, 72420S (2009).
104. We use "NICE" to generically refer to both the single-scale and multiscale implementations of NICE, and specific implementations of NICE (e.g., NICE$_{\text{Canny}}$) will be identified when necessary.
105. Using the fine-scale steerable pyramid filters to identify image contours for MS-NICE lead to statistically similar performance to the single-scale implementation of NICE using the Sobel Canny edge detectors.
106. A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," J. Opt. Soc. Am. A **14**, 2379–2391 (1997).
107. H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," IEEE Trans. Image Process. **14**, 2117–2128 (2005).
108. The subscript k for N_k accounts for decimated wavelet decompositions, such as the steerable pyramid, whose channels in coarser image scales have fewer coefficients than channels in finer image scales.
109. U. Polat and D. Sagi, "Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments," Vis. Res. **33**, 993–999 (1993).
110. U. Polat and D. Sagi, "The architecture of perceptual spatial interactions," Vis. Res. **34**, 73–78 (1994).
111. V. Kayargadde and J.-B. Martens, "Perceptual characterization of images degraded by blur and noise: experiments," J. Opt. Soc. Am. A **13**, 1166–1177 (1996).
112. D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," Proc. SPIE **6057**, 60570F (2006).
113. S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu, "Statistical edge detection: learning and evaluating edge cues," IEEE Trans. Pattern Anal. Mach. Intell. **25**, 57–74 (2003).
114. W. Ma and B. S. Manjunath, "Edgeflow: a technique for boundary detection and segmentation," IEEE Trans. Image Process. **9**, 1375–1388 (2000).
115. E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," Cogn. Psychol. **8**, 382–439 (1976).
116. C. A. Collin and P. A. McMullen, "Subordinate-level categorization relies on high spatial frequencies to a greater degree than basic-level categorization," Percept. Psychophys. **67**, 354–364 (2005).
117. C. A. Collin, "Spatial-frequency thresholds for object categorisation at basic and subordinate levels," Perception **35**, 41–52 (2006).
118. A. Torralba, "How many pixels make an image?," Vis. Neurosci. **26**, 123–131 (2009).
119. F. A. Rosell and R. H. Willson, "Recent psychophysical experiments and the display signal-to-noise ratio concept," in *Perception of Displayed Information*, L. Biberman, ed. (Plenum, 1973), pp. 167–232.
120. The Johnson criteria were based on a study with a specific set of objects, and it is possible that different objects would suggest different criteria for object recognition [11].
121. S. Ullman, *High-Level Vision: Object Recognition and Visual Cognition* (MIT, 1996).
122. S. Panis, J. De Winter, J. Vandekerckhove, and J. Wagemans, "Identification of everyday objects on the basis of fragmented outline versions," Perception **37**, 271–289 (2008).
123. D. J. Field, A. Hayes, and R. Hess, "Contour integration by the human visual system: evidence for a local "association field"," Vis. Res. **33**, 173–193 (1993).
124. The subscript k for N_k accounts for decimated wavelet decompositions, such as the steerable pyramid, whose channels in coarser image scales have fewer coefficients than channels in finer image scales.
125. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Miller, eds. (MIT, 2000), pp. 855–861.
126. M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," Appl. Comput. Harmon. Anal. **11**, 89–123 (2001).
127. B. W. Keelan, *Handbook of Image Quality: Characterization and Prediction* (CRC, 2002).
128. D. Rouse, R. Pepion, P. Le Callet, and S. Hemami, "Tradeoffs in subjective testing methods for image video quality assessment," Proc. SPIE **7527**, 75270F (2010).
129. R. W. Hamming, "Error detecting for error correcting codes," Bell Syst. Tech. J. **29**, 147–160 (1950).